How Policymakers Can Foster Algorithmic Accountability

By Joshua New and Daniel Castro | May 21, 2018

Increased automation with algorithms, particularly through the use of artificial intelligence (AI), offers opportunities for the public and private sectors to complete increasingly complex tasks with a level of productivity and effectiveness far beyond that of humans, generating substantial social and economic benefits in the process. However, many believe an increased use of algorithms will lead to a host of harms, including exacerbating existing biases and inequalities, and have therefore called for new public policies, such as establishing an independent commission to regulate algorithms or requiring companies to explain publicly how their algorithms make decisions. Unfortunately, all of these proposals would lead to less AI use, thereby hindering social and economic progress.

Algorithmic accountability: the principle that an algorithmic system should employ a variety of controls to ensure the operator can verify it acts in accordance with its intentions, as well as identify and rectify harmful outcomes.

Policymakers should reject these proposals and instead support algorithmic decision-making by promoting policies that ensure its robust development and widespread adoption. Like any new technology, there are strong incentives among both developers and adopters to improve algorithmic decision-making and ensure its applications do not contain flaws, such as bias, that reduce their effectiveness. Thus, rather than establish a master regulatory framework for all algorithms, policymakers should do what they have always done with regard to technology regulation: enact regulation only where it is required, targeting specific harms in particular application areas through dedicated regulatory bodies that are already charged with oversight of that particular sector. To accomplish this, regulators should pursue algorithmic accountability—the principle that an algorithmic system should employ a variety of controls to ensure the operator (i.e., the party responsible for deploying the algorithm)

can verify it acts in accordance with its intentions, as well as identify and rectify harmful outcomes. Adopting this framework would both promote the vast benefits of algorithmic decision-making and minimize harmful outcomes, while also ensuring laws that apply to human decisions can be effectively applied to algorithmic decisions.

INTRODUCTION

The 1990s and early 2000s saw the rise of the Internet economy, as firms used a new global network to innovate. Then, over the last decade, the world moved into the data economy, as firms increasingly used data to drive improvements in products and services. Today, the global economy is changing once again with the rise of the algorithmic economy, in which many organizations' success directly correlates with their ability to automate processes using artificial intelligence (AI).

Each of these periods has spawned many new policy questions, the answers to which have been integral to the success of digital innovation. The United States has been a leader in the digital economy, in both the development and use of information technology. It can credit its success in large part to its policies having largely rejected the precautionary principle—the idea that innovations must be proven safe before they are deployed—and the notion that the government's role is to be a speed bump—or even road block—to technological progress. Instead, the United States has embraced the innovation principle—the idea that the majority of innovations overwhelmingly benefit society, and the government's role should be to pave the way for widespread innovation while building guardrails, where necessary, to ensure public safety.

In the Internet economy, securing this success involved rules such the Internet Tax Freedom Act, which prohibits federal, state, and local governments from taxing Internet access and imposing multiple and discriminatory taxes on e-commerce, and Section 230 of the Communications Decency Act (CDA), which prevents Internet service providers from being held liable for the criminal activity of their users.[1] In the data economy, this meant avoiding comprehensive data-protection rules that limit data sharing and reuse, and instead focusing on developing tailored regulations for specific sectors, thereby allowing most industries the freedom to innovate. These policies formed the core regulatory environment that allowed companies from Amazon and eBay to Google and Facebook to thrive, and provided a distinct alternative to the precautionary, innovation-limiting rules Europe adopted.

Today, as the data economy transforms into the algorithmic economy, there is a growing chorus of voices calling for nations around the world to apply precautionary-principle regulations to the algorithmic economy. Many advocacy groups are calling for the United State to, among other things,

disregard its historic light-touch approach and mirror the policies of the European Union, especially the newly enacted General Data Protection Regulation (GDPR), which as the Center for Data Innovation has shown, is poised to relegate Europe to second-tier status in the use of AI.[2] It would be a mistake for the United States, as well as other nations wishing to replicate the United States' unprecedented success in the digital economy, to go down this precautionary-principle path.

The calls for restrictive regulation of algorithms, particularly AI, stem from widespread but incorrect beliefs that there is something inherently suspect about the technology, organizations will have strong incentives to use the technology in ways that harm individuals, and existing laws are insufficient to effectively oversee the use of this technology. Indeed, fears that algorithms could exhibit and exacerbate human bias, including facilitating discrimination and exploitation, have dominated discussions about how policymakers and regulators should treat algorithmic decision-making.[3] High-profile stories about the potential harms of algorithms, such as risk-assessment algorithms in the criminal justice system that exhibit racial bias, or advertising algorithms that promote high-paying job opportunities to men more than women, demonstrate the high stakes posed by certain algorithmic decision-making.[4] But the likelihood of these risks coming to fruition is often overstated, as advocates incorrectly assume market forces would not prevent early errors or flawed systems from reaching widespread deployment. Moreover, the solutions proposed thus far are inadequate. Some limit innovation, such as by prohibiting the use of algorithms that cannot explain their decision-making—despite being more accurate than those that can. Others fail to adequately prevent consumer harm while also limiting innovation, such as by mandating businesses disclose the source code of their algorithms, which would not effectively protect consumers and raises intellectual property concerns.

Fortunately, policymakers have an alternative to these flawed approaches. Instead of pursuing heavy-handed regulations or ignoring these risks, they should adopt the tried-and-true approach of emphasizing light-touch regulation, with tailored rules for certain regulated sectors that fosters the growth of the algorithmic economy while minimizing potential harms. The challenge for regulators stems from the fact that innovation, by its very nature, involves risks and mistakes—the very things regulators inherently want to avoid. Yet, from a societal perspective, there is a significant difference between mistakes that harm consumers due to maleficence, negligence, willful neglect, or ineptitude on the part of the company, and those that harm consumers as a result of a company striving to innovate and benefit society. Likewise, there should be a distinction between a company's actions that violate regulations and cause significant harm to consumers or competitors, and those that cause little or no harm. If

regulators apply the same kind of blanket penalties regardless of intent or harm, the result will be less innovation.[5]

To achieve a balance, regulators should take a harms-based approach to protecting individuals, using a sliding scale of enforcement actions against companies that cause harm through their use of algorithms, with unintentional and harmless actions eliciting little or no penalty while intentional and harmful actions are punished more severely. Regulators should focus their oversight on operators, the parties responsible for deploying algorithms, rather than developers, because operators make the most important decisions about how their algorithms impact society.

This oversight should be built around algorithmic accountability—the principle that an algorithmic system should employ a variety of controls to ensure the operator can verify algorithms work in accordance with its intentions and identify and rectify harmful outcomes.

When an algorithm causes harm, regulators should use the principle of algorithmic accountability to evaluate whether the operator can demonstrate that, in deploying the algorithm, the operator was not acting with intent to harm or with negligence, and to determine if an operator acted responsibly in its efforts to minimize harms from the use of its algorithm. This assessment should guide their determination of whether, and to what degree, the algorithm's operator should be sanctioned. Defining algorithmic accountability in this way also gives operators an incentive to protect consumers from harm and the flexibility to manage their regulatory risk exposure without hampering their ability to innovate.

This approach would effectively guard against algorithms producing harmful outcomes, without subjecting the public- and private-sector organizations that use the algorithms to overly burdensome regulations that limit the benefits algorithms can offer.

ALGORITHMS POSE NEW CHALLENGES

Algorithms have the potential to generate a wide variety of social and economic benefits, ranging from helping researchers increase participation in clinical trials to flagging signs of human trafficking on the deep web.[6] And with the proliferation of AI, algorithms can perform increasingly complex tasks to help solve newer and bigger challenges in the public and private sectors far more efficiently—and sometimes more effectively—than humans. However, this unprecedented complexity and scalability has also led to fears of algorithms potentially creating substantial risks that existing laws may not be able to effectively address.

COMPLEXITY

The most common criticism of algorithmic decision-making is that it is a "black box" of extraordinarily complex underlying decision models involving millions of data points and thousands of lines of code. Moreover, the model can change over time, particularly when using machine learning algorithms that adjust the model as the algorithm encounters new data. Further complicating things, in many cases, developers lack the ability to precisely explain how their algorithms make decisions, and instead can only express the degree of confidence they have in the accuracy of the algorithms' decisions.[7] The difficulty arises from the fact that while developers or operators can control what data goes into their systems, and instruct algorithms how to weigh different variables, it can be challenging, if not impossible, to program their systems to explain or justify their decisions.[8] As a result, many have labeled these algorithms as impenetrable black boxes that defy scrutiny.[9]

Complexity can be problematic for several reasons. First and foremost, it creates opportunities for bias to inadvertently influence algorithms in a number of different ways. The data algorithms train on can be flawed, such as reflecting historical biases or being incomplete, which developers or operators could fail to account for.[10] For example, if a university inadvertently denies admissions to a particular demographic at an unfair rate relative to other demographics, and then trains an algorithm to make admissions decisions based on historical admissions data, the algorithm could interpret this bias as a relevant decision-making parameter. Similarly, if developers were to train a facial-recognition algorithm on a dataset that consists primarily of images of white men's faces, it may not be able to accurately recognize faces of black women.[11] Additionally, algorithmic systems could be subject to feedback loops that perpetuate and amplify biases over time.[12] For example, consider a court system that routinely sentences blacks more harshly than whites for the same crime.[13] If that court were to implement a decision support system for sentencing that used machine learning and historical sentencing data to inform judges' decisions, that system could recommend harsher sentences for blacks based on the examples it learned from. Over time, this could serve as confirmation for a judge's unconscious bias and thus exaggerate sentencing disparities along racial lines—which can lead to increased recidivism rates and subject more blacks to harsher sentences.[14] Compounding all of this, the lack of diversity in the developer community creates the risk of homogenous developer teams failing to consider how their own unconscious biases may influence their work, such as not recognizing their training data as not being representative.[15] It should be clear, however, that in almost all of these cases the outcomes are avoidable, as developers can account for these risks and control for bias in their algorithms.

In other cases, the complexity of algorithms causes some to fear that corporations or governments could hide behind their algorithm and use algorithmic decision-making as a cover to deliberately exploit, discriminate, or otherwise act unethically.[16] For example, in her book *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor*, author Virginia Eubanks describes how policymakers in Indiana decided to implement an automated system for determining welfare eligibility.[17] While the stated goal of the switch was to increase efficiency and combat fraud, the lack of evidence regarding substantial amounts of fraud in the original system, combined with the dramatic increase in erroneous benefits denials after transitioning to the automated system, led Eubanks to conclude the system had been deliberately designed to covertly cut welfare spending without the need to change policy.[18] Eubanks fears that the public sector could exploit the use of algorithms to "avoid some of the most pressing moral and political challenges of our time—specifically poverty and racism."[19] Some also worry that algorithms could be used as covers for negligence. For example, a 2017 ProPublica investigation revealed that Facebook's advertising algorithm could allow advertisers to target anti-Semitic users by automatically generating categories of users to target for ads based on topics the users liked, which included "Jew hater" and "History of 'why jews ruin the world.'"[20] Dave Lee of the *BBC* contends Facebook tried to deflect responsibility for this by faulting their algorithm, rather than owning up to a lack of oversight—although Lee offers no evidence of Facebook intentionally trying to court anti-Semitic advertisers.[21] What is more likely, however, is Facebook's system automatically pulled data about users' likes, which in some select cases included bigoted views.

Requiring the rollout of every new technology to go perfectly would doom most of it to the scrap heap of history. All new technologies improve over time; as society interacts with them and identifies problems, developers improve the technology. Granted, this does not necessarily prevent organizations from denying responsibility for any misuse of their algorithm. But requiring an error rate of zero would considerably stifle innovation.

SCALABILITY

Another aspect of algorithmic decision-making that poses a challenge is its capacity to make a large number of decisions significantly faster than humans. As the public and private sectors increasingly rely on algorithms in high-impact sectors such as consumer finance and criminal justice, a flawed algorithm could potentially cause harm at higher rates. As existing legal oversight may not be sufficient to respond quickly or effectively enough to mitigate this risk, it is clear why increased risk warrants greater regulatory scrutiny.

By automating human-led processes, such as determining loan eligibility, banks could use algorithms to dramatically shorten the time it takes to evaluate applicants while reducing operating costs, and then pass those savings on to borrowers in the form of lower interest rates. However, if these algorithms are flawed, the sheer volume of their decisions could end up significantly amplifying the potential negative impact of these flaws. Compared with a single human, whose output is only a handful of loan applications per week, routinely making errors while evaluating loan applications, a flawed algorithm misevaluating hundreds of loan applications per week across an entire bank branch would clearly cause harm at a much larger scale.

In most cases, flawed algorithms hurt the organization using them. Therefore, organizations have strong incentives to not use biased or otherwise flawed algorithmic decision-making and regulators are unlikely to need to intervene. For example, banks making loans would be motivated to ensure their algorithms are not biased because, by definition, errors such as granting a loan to someone who should not receive one, or not granting a loan to someone who is qualified, costs banks money. But in other cases, where the cost of the error falls largely on the subject of the algorithmic decision, these incentives may not exist. Biased algorithms in parole decision systems, for instance, hurt individuals who are unfairly denied parole, but impose little cost on the court system. In such cases, existing legal frameworks may not be sufficiently equipped to respond quickly or effectively to mitigate this risk.

Of course, if an organization has a flawed process for human decisions, the impact could also be significant—such as when banks changed their lending practices to extend credit to borrowers who had little or no documentation of income contributing to the 2008 financial crisis.[22]

FLAWED REGULATORY PROPOSALS

While there has been a growing call for policymakers to mitigate the risks of algorithmic decision-making, proposed solutions are typically ineffective, counterproductive, or harmful to innovation. These proposals fall into three main categories: calls for algorithmic transparency, explainability, or both; calls for the creation of regulatory bodies to oversee all algorithmic decision-making; and generalized regulatory proposals, or proposals that rely so heavily on poorly articulated or vague concepts that they are simply not viable. Most of these proposals endorse the precautionary principle and are based on the belief that algorithms, particularly AI, should be proactively regulated and proven safe before being deployed—and once deployed, should be heavily regulated. But there are also a number of people who believe government should not regulate emerging technologies and should leave industry solely responsible for addressing the potential

harms of algorithmic decision-making. While many types of algorithmic decision-making do not require additional regulatory oversight, some do.

It is important to note that certain aspects of these proposals do have merit, and some of these concepts are valid and useful components of algorithmic accountability. However, while they have their place in particular contexts, it would be inappropriate to apply these policies across all sectors of the economy.

ALGORITHMIC TRANSPARENCY AND EXPLAINABILITY

The most common proposal for regulating algorithms focuses on the principle of algorithmic transparency, which requires organizations to expose their algorithms and information about their data to some degree of public scrutiny. Supporters define this principle in different ways, but the common theme is algorithmic transparency is based on the notion that the complexity and proprietary nature of algorithms can obscure how they make decisions and thus mask harmful behavior. Algorithmic transparency advocates believe that exposing the code and underlying data of these black boxes would allow the public and regulators to identify whether and how an algorithm is producing harmful outcomes.

Support for algorithmic transparency is widespread, both in the United States and abroad.[23] A Pew survey found that many technologists believe algorithmic transparency would be a good way to mitigate the risks of algorithms, while the U.S. Federal Trade Commission (FTC) has expressed support for algorithmic transparency—though it is unclear exactly how the FTC defines it.[24] Cathy O'Neil, author of the book *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* and founder of O'Neil Risk Consulting & Algorithmic Auditing (ORCAA), a consultancy that helps companies manage algorithmic risks, writes, "Models that have a significant impact on our lives, including credit scores and e-scores, should be open and available to the public," and that certain potentially harmful algorithms "must also deliver transparency, disclosing the input data they're using as well as the results of their targeting."[25] Additionally, the Electronic Privacy Information Center (EPIC) states, "Algorithmic transparency should be established as a fundamental requirement for all AI-based decision-making."[26] EPIC would have regulators go even further, asserting, "The algorithms employed in big data should be made available to the public."[27] In effect, many in this camp believe any computer system that uses automated decisions should make its source code available for some degree of public scrutiny.

At the same time, there are others who call for more general transparency. Market research executive Barry Chudakov thinks companies should include the equivalent of a nutrition label for their algorithms, indicating how an algorithm might make certain decisions, and the implications of

those decisions.[28] Ben Wagner, director of the Centre for Internet and Human Rights, argues that companies should disclose whether decisions they make on their platforms are made by algorithms or humans.[29]

Others, such as Judith Donath, a fellow at Harvard University's Berkman Klein Center for Internet & Society, lament the opacity of complex algorithmic systems, arguing:

> The danger in increased reliance on algorithms is that the decision-making process becomes oracular: opaque yet unarguable. The solution is design. The process should not be a black box into which we feed data and out comes an answer, but a transparent process designed not just to produce a result, but to explain how it came up with that result. The systems should be able to produce clear, legible text and graphics that help the users—readers, editors, doctors, patients, loan applicants, voters, etc.—understand how the decision was made.[30]

While various players in this camp state they want "transparency," they typically mean "explainability," which are two commonly conflated terms in discussions about governing algorithms.[31] Transparency refers to disclosing an algorithm's code or data (or both), while explainability refers to the concept of making algorithms interpretable to end users, such as by having operators describe how algorithms work or by using algorithms capable of articulating the rationales for their decisions. For example, the European Union has made explainability a primary check on the potential harms of algorithmic decision-making, guaranteeing in its GDPR the right for a person to obtain "meaningful information" about certain decisions made by an algorithm.[32] Similarly, France's Secretary of State for Digital Affairs, Mounir Mahjoubi, has stated that the government should not use an algorithm if it cannot explain its decisions.[33]

While transparency and explainability are fundamentally different concepts, they share many of the same flaws as a solution for regulating algorithms. First, they hold algorithmic decisions to a standard that simply does not exist for human decisions. As EPIC describes, "Without knowledge of the factors that provide the basis for decisions, it is impossible to know whether government and companies engage in practices that are deceptive, discriminatory, or unethical. Therefore, algorithmic transparency is crucial to defending human rights and democracy online."[34] This argument fails to recognize that algorithms are simply a recipe for decision-making. If proponents of algorithmic transparency and explainability are concerned that these decisions are harmful, then it is counterproductive to only call for algorithmic decisions to be transparent or explainable, rather than for all aspects of all decision-making to be made public or explained. If blanket mandates for transparency and explainability are appropriate for algorithmic decision-making, but not human decision-making (which itself is often supported by computers), logic would dictate that human decisions

are already transparent, fair, and free from unconscious and overt biases. In reality, bias permeates every aspect of human decision-making, so to hold algorithms to a higher standard than for humans is simply unreasonable. For example, research shows taxicabs frequently do not pick up passengers based on their race, and employers may eliminate job applicants with African-American sounding names despite their sufficient qualifications.[35] Yet, understandably, taxi drivers are not required to publicly report their reasons for not picking up every passenger they pass by, and employers do not have to publish a review of every resume they receive, with detailed notes explaining why they choose not to offer a particular candidate a job, because laws and regulations for these sectors focus on outcomes, not unconscious bias. If EPIC and other proponents of algorithmic transparency and explainability worry that such broad categories of decisions have the potential to be harmful due to the influence of bias, then they should advocate for transparency and explainability in all significant decision-making, as an algorithm's involvement in those decisions is irrelevant.

Second, calls for the right to meaningful information about certain algorithmic decisions, as the European Union's GDPR mandates, disregard the many laws that already exist guaranteeing a right to an explanation for certain high-impact decisions, such as the reasons behind a bank refusing to grant an applicant a loan, or why a company fired an employee.[36] Existing laws would still apply to these situations regardless of whether companies use an algorithm to make the decision. In application areas where laws already exist, new requirements specifically targeting algorithms would be redundant—although the GDPR extends this requirement to all algorithmic decisions with legal or significant consequences.[37] If there are certain decisions that warrant an explanation or meaningful information, then surely it should not matter whether an algorithm was involved. But if these decisions do indeed carry potential risks, the construction of this requirement allows for companies to use humans instead of algorithms to skirt the law.[38] If the GDPR's supporters believe such decisions warrant an explanation, then it is ineffective for the GDPR to only target decisions made by algorithms.

Proponents for algorithmic transparency often justify their stance by pointing to the potential for biased and flawed algorithms in the criminal justice system to cause substantial harm to individuals. As this paper discusses, transparency, as well as other components of algorithmic accountability such as error analysis and procedural regularity, will likely be key factors to ensure the beneficial use of algorithms wherever market forces are muted, such as with the criminal justice system. However, the value of transparency in the criminal justice context does not support the conclusion that algorithmic transparency would be necessary or beneficial in most contexts. As noted above, for most applications, operators have

strong incentives to minimize flaws and potential harms. But for applications that lack these incentives, whether an operator uses algorithms is irrelevant. It is also important to bear in mind that even in the criminal justice system, algorithmic transparency would not address the root causes of many of the harms that such decisions can cause. For example, algorithmic transparency alone would not solve inherent bias problems, such as the large disparity in arrest rates for blacks and whites for marijuana possession, despite marijuana use being roughly equal among blacks and whites.[39]

Third, another major flaw with more extreme demands for transparency, such as EPIC's call for all source code to be made fully public, is that while "pulling back the curtain" to allow regulators and the public to scrutinize how an algorithm might be flawed may sound reasonable, it is unrealistic to expect that even the most technologically savvy, resource-flush regulators, advocacy groups, or concerned citizens would be capable of reliably gleaning meaningful information from scrutinizing advanced AI systems and their underlying data, particularly at scale. For example, after Reddit disclosed a portion of its ranking algorithm, a group of computer scientists led by Christian Sandvig at the University of Michigan noted that "even with complete transparency about a particular part of [this] algorithm, expert programmers have been sharply and publicly divided about what exactly that part of the algorithm does. This clearly implies that knowing the algorithm itself may not get us very far in detecting algorithmic misbehavior."[40] While examining code can provide meaningful information about how some algorithmic systems make decisions, for many advanced AI systems that rely on thousands of layers of simulated neurons to interpret data, even their developers cannot explain their decision-making. For example, researchers at Mount Sinai Hospital in New York developed an AI system called Deep Patient that can predict whether a patient is contracting any of a wide variety of diseases.[41] The researchers trained Deep Patient on the health data from 700,000 patients, including hundreds of variables, which allow it to predict disease without explicitly having to be taught how.[42] The system is substantially better than other disease-prediction methods, yet its own developers do not know how its decision-making process works.[43] Thus, there is little to reason to believe a third party would be able to understand it. As Curt Levey of the Committee for Justice and Ryan Hagemann of the Niskanen Center describe, "The machine's 'thought process' is not explicitly described in the weights, computer code, or anywhere else. Instead, it is subtly encoded in the interplay between the weights and the neural network's architecture. Transparency sounds nice, but it's not necessarily helpful, and may be harmful."[44] The United Kingdom's Government Office for Science cautions, "Most fundamentally, transparency may not provide the proof sought: Simply sharing static code provides no assurance it was actually used in a

particular decision, or that it behaves in the wild in the way its programmers expect on a given dataset."[45]

Fourth, calls for algorithmic transparency and, sometimes, for algorithmic explainability discount the value of proprietary software. Requirements to publicly disclose source code or information about the inner workings of software would reduce incentives for a company to invest in developing algorithms, as competitors could simply copy them. While copyright laws could reduce this risk in countries with strong intellectual property protections like the United States, this would make it significantly easier for bad actors in countries that routinely flout intellectual property protections, such as China, to steal source code.[46] Ardent supporters of algorithmic transparency, such as Frank Pasquale, author of *The Black Box Society: The Secret Algorithms That Control Money and Information*, dismiss this concern out of hand, claiming the argument is just a nefarious smoke screen to cover for deliberate exploitation or abuse: "They [corporations] say they keep techniques strictly secret to preserve valuable intellectual property—but their darker motives are also obvious."[47] It is not clear what these darker motives are, other than to maximize profits. But again, in almost all cases where companies stand to lose from an algorithmic system making biased decisions, the company is highly motivated to make accurate decisions—unless Pasquale is arguing that accurate decisions, like denying a loan to someone who presents a bad credit risk is a reflection of a "darker" motive, it is hard to know what the problem is.

Fifth, requiring algorithmic transparency can also create opportunities for bad actors to "game the system" and take advantage of algorithm-driven platforms. For example, for years, Google relied on an algorithm called PageRank to determine the order of search results to display based on factors such as a website's meta tags and keywords.[48] However, because these factors were widely known, any site owner could manipulate the algorithm by populating a page with hidden content that PageRank interpreted as desirable in an effort to push their website higher in the search rankings and increase views, despite it being irrelevant to a user's query.[49] Now, Google uses a combination of multiple, complex algorithms, including machine learning systems, that weighs hundreds of factors to order search results based on content quality and relevance.[50] If Google or other search engines were required to disclose how their search algorithms work, it would once again allow websites to exploit these systems—and consumers would suffer for it.

Sixth, transparency would not solve some of the key challenges in the information economy. Some, such as German Chancellor Angela Merkel, argue, "Algorithms, when they are not transparent, can lead to a distortion of our perception," contributing to the formation of filter bubbles in online platforms (i.e., situations in which users are only exposed to content that

conforms to their world view) and damages public discourse.[51] Similarly, German Federal Minister of Justice Katarina Barley believes such practices contribute to the proliferation of online disinformation campaigns.[52] Both Merkel and Barley claim that algorithmic transparency would alleviate these problems by enabling users to understand how their perspectives are being influenced. However, it would likely have the opposite effect, making it easier for bad actors to game these algorithms and flood platforms with low-quality or deliberately misleading content.

Seventh, mandating algorithmic explainability could severely limit the potential benefits of algorithms, as there can be inescapable trade-offs between the interpretability or explainability of an AI system and its accuracy. As data scientists Max Kuhn and Kjell Johnson put it in their book *Applied Predictive Modeling*, "Unfortunately, the predictive models that are most powerful are usually the least interpretable."[53] An algorithm's accuracy typically increases with its complexity, but the more complex an algorithm is, the more difficult it is to explain.[54] While this could change in the future as research into explainable AI matures, at least in the short term, requirements for explainability would only be desirable in situations where it is appropriate to sacrifice accuracy—and these cases are rare. For example, it would not be desirable to prioritize explainability over accuracy in autonomous vehicles, as even slight reductions in navigation accuracy or to a vehicle's ability to differentiate between a pedestrian on the road and a picture of a person on a billboard could be enormously dangerous. Thus, a mandate for algorithmic explainability is essentially a mandate to use less-effective AI, or, in cases where sacrifices in accuracy are prohibitive, such as with self-driving cars, a ban on the use of effective but uninterpretable algorithms.

Finally, the most fundamental flaw in these proposals is that algorithmic transparency and explainability are a means for achieving the goal of preventing algorithms from causing harm, not an end themselves. Transparency and explainability can indeed be useful mechanisms for achieving this goal, but only in select contexts. It would be unwise for regulators to treat achieving algorithmic transparency or explainability as either a panacea or an end-goal. In most contexts, mandating transparency and explainability would limit innovation and fail to prevent potential harm.

MASTER REGULATORY BODIES TO OVERSEE ALL ALGORITHMIC DECISION-MAKING

As concerns about the potential risks of algorithms proliferate, some have advocated for governments to create new regulatory bodies specifically devoted to overseeing algorithms. For example, University of Maryland computer science professor Ben Shneiderman, in a 2017 speech at the Alan Turing Institute, proposed the creation of a "National Algorithm Safety Board" to independently oversee the use of algorithms, such as by

auditing, monitoring, and licensing algorithms when a company wants to deploy one.[55] Similarly, the Oxford Internet Institute calls for the creation of an "algorithmic oversight institution" with the powers to audit algorithms and determine whether they serve the public interest.[56] Attorney Andrew Tutt proposes the creation of the equivalent of the U.S. Food and Drug Administration for algorithms, which would have the power to "prevent the introduction of algorithms into the market until their safety and efficacy has been proven through evidence-based premarket trials."[57] Entrepreneur Elon Musk, speaking at a 2017 meeting of the National Governors Association, urged policymakers to take a precautionary principled approach, arguing that "the right order of business would be to set up a regulatory agency [with the] initial goal: gain insight into the status of AI activity, make sure the situation is understood, and once it is, put regulations in place to ensure public safety."[58]

These and related proposals suffer from a number of serious challenges. First, they all fail to recognize that to adequately assess an algorithmic decision, one would need to have context-specific knowledge about the type of decisions an algorithm is dealing with. What constitutes harm in consumer finance involves dramatically different criteria than what constitutes harm in health care, which is why governments have different sector-specific regulatory bodies. If it would be ill-advised to have one government agency regulate all human decision-making, then it would be equally ill-advised to have one agency regulate all algorithmic decision-making. This is why during the growth of the Internet in the 1990s, the United States did not establish a federal Internet agency to regulate all online activity, as some proposed. Instead, the Federal Communications Commission regulated the telecommunications aspects of the technology, the FTC regulated online commerce, the National Telecommunications and Information Administration regulated spectrum, and so on.

It is unclear why advocates for these proposals believe existing regulatory bodies are incapable of scrutinizing algorithms effectively. It is important for regulators to understand the technology related to issues under their purview and, given the newness of this technology, it is likely that many agencies lack the technical expertise to understand how algorithmic decision-making works. For example, after Congress proposed the U.S. National Highway Transportation Safety Administration (NHTSA) be responsible for certifying the safety of autonomous vehicles, Mike Ramsay, an automotive technology analyst at Gartner Research, lamented that "there's no way NHTSA has the technical capability to do this right now."[59] However, agencies often lag behind the private sector in their ability to understand new technologies, and have always had to deal with the issue of staying informed about new innovations. Moreover, some agencies, such as the FTC, actively cultivate and seek out technical expertise to allow them to effectively oversee complicated technology issues in a wide array

of industries.[60] Regardless, if the concern is government agencies not having sufficient technical expertise, simply establishing a new regulatory agency devoted to algorithms would not fix this, as any difficulties governments face in attracting and retaining human capital would still apply.[61]

This does not mean Congress and other legislative bodies should not support agencies in developing the needed technical expertise to manage AI-related concerns. Stanford University's One Hundred Year Study on Artificial Intelligence, or AI100, led by a group of academics and AI experts, recommends that policymakers:

> Define a path toward accruing technical expertise in AI at all levels of government. Effective governance requires more experts who understand and can analyze the interactions between AI technologies, programmatic objectives, and overall societal values. … Absent sufficient technical expertise to assess safety or other metrics, national or local officials may refuse to permit a potentially promising application. Or insufficiently trained officials may simply take the word of industry technologists and green light a sensitive application that has not been adequately vetted. Without an understanding of how AI systems interact with human behavior and societal values, officials will be poorly positioned to evaluate the impact of AI on programmatic objectives. … Faced with the profound changes that AI technologies can produce, pressure for "more" and "tougher" regulation is inevitable. Misunderstanding about what AI is and is not, especially against a background of scare-mongering, could fuel opposition to technologies that could benefit everyone. This would be a tragic mistake.[62]

This is not to say regulatory regimes should never change as algorithmic decision-making proliferates and AI matures—although every regulator modernizes in tandem with its sector. The U.S. Federal Aviation Administration will likely operate substantially differently in 50 years if flying cars become commonplace, just as the European Medicines Agency will have to adapt should cancer treatments that rely on nanorobotics become the norm. But reworking a government's entire regulatory system in response to just a single technology would be a dramatic and likely ineffective measure.

Establishing a regulator to oversee the use of algorithms also implies that all algorithms pose the same level of risk and need for regulatory oversight. However, algorithms pose a wide variety of risk depending on their application. Low-risk decisions should not be subject to regulatory oversight simply because they use an algorithm.

Finally, some of these proposals focus on serving the "public interest," even though reasonable people can differ on what this should include. Use of personal vehicles, for example, could be considered in the public

interest because they provide mobility and access to economic opportunity—although they also pollute the environment, create sprawl, and kill upwards of 30,000 people a year in the United States. Rather than have a regulator decide whether use of vehicles is in the public interest, government instead regulates specific activities based on their objective benefits and harms, such as the fuel economy and safety of vehicles and land use in cities. Similarly, policymakers should not decide whether the use of algorithms are in the public interest, but instead regulate specific uses of them.

GENERALIZED REGULATORY PROPOSALS

The third category of proposals is a disparate group of likely well-intentioned recommendations that are vague and memorable, but ultimately just meaningless slogans and buzzwords that are unworkable from a regulatory perspective. To be sure, some of these proposals could have value in certain areas of algorithmic decision-making, such as in AI development guidelines or corporate social responsibility standards, but as guides for regulation they are impractical.

There are countless examples of calls to action for regulating algorithms that stress the need to rethink current approaches but fail to articulate an effective path forward. Emblematic is a January 2018 speech from British Prime Minister Theresa May, who stated that while the potential of AI is fundamental to the advancement of humanity, "This technological progress also raises new and profound challenges which we need to address. … So today I am going to make the case for how we can best harness the huge potential of technology. But also how we address these profound concerns."[63] However, the rest of the speech failed to actually propose any meaningful solutions to potential challenges posed by AI. As one critic writes, May's argument can be boiled down to "AI can do great things, but we must be sure it's safe and ethical," and that this is "vapid, a truism."[64]

In some cases, these narratives are presented as guiding principles to help policymaking rather than as bona fide policy proposals themselves. While they are designed to serve as a reference for future efforts, they often include recommendations that are either too vague to be useful or that would restrict beneficial uses of algorithms. For example, a March 2018 report from the European Group on Ethics in Science and New Technologies, which advises the European Commission, that called for the creation of a shared ethical and regulatory framework for AI concluded, "The principle of responsibility must be fundamental to AI research and application. 'Autonomous' systems should only be developed and used in ways that serve the global social and environmental good, as determined by the outcomes of deliberative democratic processes."[65] The report's attempt to explain this is even more vague, stating, "[Autonomous systems] should be designed so that their effects align with a plurality of

fundamental human values and rights."[66] While this may sound innocuous, it is incredibly problematic. Would it be acceptable to deploy algorithms to deliberately facilitate discrimination in societies where the plurality of human values is to limit rights for women or religious minorities? This lack of specificity gives policymakers little to work with when it comes to crafting regulation. The problem with such proposals is that they do not specify who decides what values to endorse or how to reconcile trade-offs, such as between job loss and economic growth. Additionally, this approach could potentially prohibit the use of algorithms for the purpose of increasing productivity, with no impact on human rights.

More importantly, such assertions ignore that democratic societies have processes by which they adjudicate these conflicting interests and values: legislatures and courts of law. Yet some in the AI community seem to think they are self-appointed guardians of ethics, as they define it. For example, some have argued against autonomous weapons systems, arguing developers should only create algorithms for robots that augment—but do not replace—human workers. While the intent of these proposals is to save lives and jobs, decisions like these should ideally be made through democratic processes, not by a select group of individuals who may not reflect the broad diversity of society. Nation states, for instance, are best suited to determine what defense systems they need to protect themselves from adversaries. Furthermore, social and political preferences are normally not applied to technologies, but rather to specific sectors and industries. Because of these constraints, and the need to satisfy consumers, firms are better suited to determine how to maximize innovation.

Some of these proposals attempt to take a more productive approach but are still ultimately unworkable. For example, in May 2016, the White House published a report detailing the opportunities and challenges of big data and civil rights. But rather than focus on demonizing the complex and necessarily proprietary nature of algorithmic systems, it presented the concept of "equal opportunity by design," which it defined as the principle of ensuring fairness and safeguarding against discrimination throughout a data-driven system's entire lifespan.[67] This approach, described more generally by then Federal Trade Commissioner Terrell McSweeny as "responsibility by design," recognizes that algorithmic systems can produce unintended outcomes, and encourages developers to address the root problems that could cause harms in algorithmic systems, such as failing to account for historical bias.[68] Encouraging developers to be responsible in the creation and application of algorithms is a worthwhile goal, however merely stating developers should consider "responsibility by design" is not a clear solution to the challenges algorithms pose.[69] Furthermore, these approaches focus on the developer, not the operator. While developers could wholeheartedly embrace "responsibility by design," it would have

little impact if their algorithms were not viable products. Rather, if operators were exposed to regulatory incentives to deploy algorithms responsibly, the market would respond to this demand much more efficiently.

DOING NOTHING IS NOT THE ANSWER

Overzealous regulation of technology, inspired by fears of worst-case scenarios or beliefs that a select group of AI developers are the only ones that can know and protect consumer interests will clearly harm innovation. Speculative fears about the potential risks of new technology are a powerful driver of advocacy efforts to restrict how a technology can be used before it matures to the point where society can fully realize its benefits and understand its impacts. Fears that preempt the proliferation of disruptive technologies have spurred regulatory proposals that seem ridiculous in retrospect after the technology becomes commonplace.[70] For example, as transistors proliferated in the 1950s and 1960s, some U.S. policymakers were so concerned about their potential to be used for surveillance that one senator proposed a law that would have required all bugging equipment to be licensed by the government.[71] Had Congress succumbed to such hysterical concerns and passed that bill, many innocuous technologies that are widely enjoyed today, such as smartphones and baby monitors, would have been greatly impeded.

Thus, it is wise to be skeptical of advocates rushing to regulate new technologies due to concerns about their hypothetical harms before it is clear how market forces, technological advancement, and existing regulations would shape their use as they mature. However, with algorithmic decision-making, dismissing any and all efforts to improve governance would be problematic. While explicit calls for the government to not regulate any algorithms and leave it entirely to industry to self-regulate are few and far between, some do advocate for it. For example, technology reporter Tristan Greene, writing for *The Next Web*, concluded that due to the speculative nature of many of the fears about AI, the "government is clueless about AI and shouldn't be allowed to regulate it."[72] Mouloud Dey, director of innovation and business solutions at SAS France, argues that governments should not step in to regulate algorithms because of the burden regulations could have on innovation—and that industry self-regulation would be adequate to address any potential harms.[73] In many cases however, more general anti-regulation attitudes could still lend credence to the notion that the government should not regulate algorithms at all, by overshadowing legitimate efforts to regulate the technology in an evenhanded, beneficial way. For example, Simon Constable, a fellow at the Johns Hopkins Institute for Applied Economics, Global Health, and the Study of Business Enterprise, writing in *Forbes*, erroneously concluded that due to the U.S. government's failure to prevent or mitigate the 2008

financial crisis, “It's time to just say no to calls for more government regulation of the tech industry.”[74]

Given the steps some governments, such as the EU’s GDPR, have already taken that will clearly limit innovation, it is easy to be sympathetic to such positions. But while industry self-regulation, market forces, and tort law will likely play a large role in positively shaping the use of algorithms, there are reasons why these alone would be insufficient to protect against all potential harms of algorithmic decision-making, which likely fall into one of three categories. First, there are some potential applications of algorithms where traditional market forces that could mitigate the harms of algorithms, such as the threat of reputational damage if a company’s algorithm causes harm, are diminished, making the cost of this flawed decision-making one-sided. This is particularly true with government uses of AI wherein the costs of bad decisions are indeed problematic, but not borne directly by the government agency using the algorithm. In other words, even though a discriminatory algorithm is an inferior product, there are some situations where this would not deter an operator from deploying it. Second, there are applications of algorithmic decision-making where even though incentives to minimize harms exist, the potential harms could be significant enough to warrant regulation, such as is with autonomous vehicles. And third, certain applications of algorithms could cause harms, such as exacerbating inequality, but without an operator expressly or obviously breaking the law. For example, an online jobs board could utilize a targeted advertising algorithm that does not consider race but nonetheless uses variables that inadvertently serve as proxies for race, such as zip code, thereby favoring members of a certain race for job opportunities. This harm may not be immediately obvious to the public, regulators, or even the operator. In such cases, absent public outrage, businesses have reduced incentive to scrutinize their algorithms thoroughly to prevent this harm, as there is not a strong profit motive to do so.

Table 1: Summary of major proposals for algorithmic governance and their flaws

Proposal	Flaws
Algorithmic transparency or explainability mandates	▪ Holds algorithmic decisions to a standard that does not exist for human decisions ▪ Incentivizes organizations to not use algorithms, thus sacrificing productivity ▪ Fails to address the root cause of potential harms ▪ Assumes the public and regulators could interpret source code for complex algorithms even developers themselves cannot always understand ▪ Undermines closed-source software, reducing incentives for innovation ▪ Makes it easy for bad actors to "game the system" ▪ Creates incentives for the use of less-effective AI, as there can be trade-offs between explainability and accuracy for complex AI ▪ Is useful in select contexts but ineffective or harmful in others
Regulatory bodies to oversee all algorithmic decision-making	▪ Ignore the need for regulators to have context-specific expertise ▪ Low-risk decisions should not be subject to regulatory oversight
Generalized regulatory proposals	▪ Provide no specifics on how to operationalize proposals ▪ Rely heavily on platitudes that do not translate to effective governance
Do nothing	▪ Does not recognize that, for some use cases—particularly certain government applications—algorithms are less subject to market forces that would minimize potential harms ▪ Fails to prevent algorithms from causing harm in certain contexts where such harms are not obvious or do not break the law

DEFINING ALGORITHMIC ACCOUNTABILITY

Instead of the approaches just discussed, what is needed, at least for some applications, is algorithmic accountability. To be sure, the concept of algorithmic accountability itself is at risk of becoming a buzzword, as many have conflated it with other concepts. For example, Ashkan Soltani, former chief technologist of the FTC, said that although the FTC's stated goal is to

pursue algorithmic transparency, "maybe 'accountability' would have been a better term to use," noting that making companies turn over source code is not always an effective solution.[75] As the World Wide Web Foundation describes, "When applied to algorithms, algorithmic accountability has often been conflated with other values, such as transparency. … However, several researchers in recent years have pointed to limitations in defining algorithmic accountability as transparency. … Although we are at a stage in which the definition of algorithmic accountability is still being agreed upon, experts and practitioners have been putting forward general principles to be debated."[76]

Although some attempts to define algorithmic accountability are generally useful, they are not much help for regulators. For example, in 2016, an organization of computer scientists and researchers called the Fairness, Accountability and Transparency in Machine Learning (FAT/ML) community published its "Principles for Accountable Algorithms," which state that responsibility, explainability, accuracy, and auditability are all key components of algorithmic accountability.[77] Unlike other proposals that tend to recommend a one-size-fits-all approach, FAT/ML's principles emphasize the variety of technical solutions available to help mitigate the potential risks of algorithms, and articulate what concepts like "responsibility" might mean from a technical perspective. However, FAT/ML's principles are geared toward the developer community and thus fall short of serving as a meaningful foundation for governance.

Perhaps the best definition of algorithmic accountability comes from the World Wide Web Foundation, which describes it as "[ensuring] harms can be assessed, controlled, and redressed" in an algorithmic system.[78] As described in the next section, this is a core component of algorithmic accountability, although the World Wide Web Foundation does not establish how algorithmic accountability relates to legal or regulatory frameworks, leaving the question of how and when this should be enforced—to achieve what they call "algorithmic justice"—open-ended.[79]

This paper's definition of algorithmic accountability has three simple goals: promote desirable or beneficial outcomes; protect against undesirable, or harmful, outcomes; and ensure laws that apply to human decisions can be effectively applied to algorithmic decisions.

Regulators should use algorithmic accountability to hold the party responsible for deploying an algorithm, the algorithm's "operator," legally accountable for its actions. For example, a government agency that uses an algorithm to screen people at border crossings, or a company that deploys an AI system to vet job applications, would be considered operators, while a developer who publishes an algorithm would not. This is important because simply creating an algorithm that exhibits some kind of demographic bias, for example, does not cause others harm and should be

of no concern to regulators unless an operator applies it in a way that could cause harm, just as it is not illegal for a person to hold biases, but it is against the law for them to base certain decisions on these biases, such as deciding whom to hire.

A governance framework for algorithmic accountability is based on the principle that an algorithmic system should employ a variety of controls to ensure operators can:

- Verify it works in accordance with the operator's intentions; and
- Identify and rectify harmful outcomes.

Algorithmic accountability promotes desirable outcomes, protects against harmful ones, and ensures algorithmic decisions are subject to the same requirements as human decisions. This approach is technology neutral, granting operators flexibility to employ a variety of different technical and procedural mechanisms to achieve algorithmic accountability. Importantly, algorithmic accountability is relevant only when an application of algorithmic decision-making poses potential harms significant enough to warrant regulatory scrutiny, and not, for example, applications that only pose the risk of minor inconveniences should the algorithms involved be flawed.

OPERATORS CAN VERIFY AN ALGORITHM ACTS IN ACCORDANCE WITH THEIR INTENTIONS

The first step in achieving algorithmic accountability is determining whether algorithms are working the way their operators intended. If the answer is yes, and it is causing harm, then it is important to recognize that regulations already exist in various industries that prohibit racial discrimination, require due process, and so on. When an operator intends to cause harm, whether they use an algorithm to do so should be irrelevant. There are a variety of different technical and procedural mechanisms that can be employed, when contextually relevant, to make the determination of whether a harm is intentional. These include: transparency, explainability, confidence measures, and procedural regularity. In most cases, operators would likely have to employ a combination of several of these mechanisms in order to be confident an algorithm is acting as they intended. This is not meant to be a comprehensive list of all the ways an operator can verify an algorithm is acting as intended, as there may be methods that are only useful in niche circumstances or that have yet to be developed.

Transparency

Despite the many ways a broad mandate for algorithmic transparency would be detrimental, transparency can play a valuable role in achieving algorithmic accountability for some applications. Both the type of algorithm

in question and the context of its application play a role in whether transparency would be desirable. As previously addressed, many black-box systems that rely on machine learning can be prohibitively complex for humans—even for the developers of these systems—to interpret in a meaningful way.[80] However, transparency could allow operators to interpret how less-complex algorithms developed by third parties ensure they are functioning as intended. For example, risk-assessment algorithms, such as those used to inform sentencing decisions, may rely on many different variables in their assessments but be static and relatively straightforward, making it is easy for their operators to assess the variables involved and determine whether they are appropriate—as well as observe how a certain data point might impact a risk score because the system is hard-coded to give that variable a particular weighting.[81] In these cases, transparency would be a simple and direct way for operators to ensure could algorithms are doing what they want them to. Different degrees of transparency would be appropriate in different contexts. For example, many businesses that lack the necessary expertise may benefit from making source code available to a nondisclosure-bound third party that is proficient in conducting code audits to identify errors.

Explainability

Like transparency, explainability can be a useful tool in achieving algorithmic accountability—but only in certain contexts. In theory, having an algorithm clearly explain how it made a decision would be the most effective and direct way to ensure an algorithmic system is acting as intended—provided the explanation is verifiably correct. For that reason, learning how to make AI explainable is an active field of research, often called "XAI."[82]

However, complexity is again a limiting factor, for two reasons. First, given the trade-off between explainability and accuracy, it is rarely desirable to prioritize explainability at the expense of accuracy. Second, developing an AI system capable of explaining itself or justifying its decisions is an incredibly challenging technical feat, so much so that the U.S. Defense Advanced Research Projects Agency (DARPA) devoted $75 million in 2017 to research how it could be achieved.[83]

Confidence Measures

Confidence measures, also known as confidence intervals, are metrics that indicate how confident an algorithm is in a decision or prediction—and there are a variety of different statistical techniques that an algorithm can use to generate confidence measures.[84] Confidence measures are a relatively simple method for ensuring an algorithm is acting as intended. For example, if a bank were to deploy an AI system to monitor and prevent fraudulent transactions, using confidence measures as a threshold for different outcomes would ensure it was not unnecessarily blocking

legitimate transactions and allowing fraudulent ones to slip though. If, for example, after scrutinizing a transaction the system was only 80 percent confident the transaction was fraudulent, it could flag the case for human review rather than automatically halt it.

Procedural Regularity

Procedural regularity is the process of consistently applying an algorithm in the same manner.[85] This can be a useful way to ensure an algorithm is acting as intended because, without the ability to explain why an algorithm made a particular decision, an operator can consistently apply the algorithm and scrutinize its outputs to evaluate whether it is achieving its desired results. As described by Joshua Kroll, a systems engineer at CloudFlare, procedural regularity is useful for the oversight of algorithms in the same way it is useful in cryptography:

> [C]ryptographic techniques can let a system prove its *procedural regularity*—in other words, that the same process was applied in all cases, similar to the idea of procedural due process in the law—even without revealing what the process is or how it operated in specific cases. Procedural regularity is an important basis for further examination of computer system behaviour—without knowing whether a particular system was actually used to make some decision (vs. whether that decision was made "on a whim" and in an arbitrary way), it is very difficult to ask whether the system is being fair or complying with the law. And without direct evidence, it is extremely difficult from a technical perspective to say whether a particular system was, in fact, used in a particular case without completely recomputing the system's decisions.[86]

OPERATORS CAN IDENTIFY AND RECTIFY HARMFUL OUTCOMES

Simply taking steps to verify an algorithm is acting as intended is not enough to ensure it is not also producing harmful outcomes. Thus, an accountable algorithmic system must also allow operators to identify and minimize harmful outcomes. This is an important capability because it allows for organizations to responsibly deploy algorithms despite not being able to predict or control for every possible harmful outcome that could arise from an algorithm's decisions—which would likely be impossible and could severely limit the utility of algorithms. There are a variety of methods to accomplish this that allow operators to take meaningful steps to minimize harms. These include, but are not limited to, impact assessment, error analysis, and bias testing. Importantly, these are not simply just post hoc controls—operators can and should apply these steps throughout the entire process of developing and deploying an algorithm, and continuously employ them throughout the time an algorithm is in use.

Impact Assessment

Just as policymakers use impact assessments to gather evidence about the potential social or economic impact of certain policies, organizations

can carry out impact assessments on particular algorithms.[87] New York University's AI Now Institute has proposed a preliminary framework for what it calls Algorithmic Impact Assessments (AIAs) for the use of algorithms in key public-sector applications, emphasizing key steps an agency can take to identify when and how potential harms could arise, including by increasing its expertise and capacity to effectively implement and evaluate algorithmic systems, and allowing third parties to audit public-sector algorithms.[88] Impact assessments are highly contextual, and what is appropriate in one domain may not be effective or desirable in another. Also, like policy impact assessments, impact assessments for algorithms can be both ex ante, focusing on prospective analysis, and ex post, focusing on continuous and historical analysis.[89] For example, the Department of Veterans Affairs could conduct an impact assessment for a mental health diagnostic algorithm to determine whether it would be appropriate to implement, and then conduct, another assessment of the algorithm after a year of use to evaluate its effectiveness; or the Department of Housing and Urban Development could conduct a disparate impact analysis after it deploys a new algorithm that could influence housing decisions.[90] Particularly sensitive applications, such as the use of algorithms in the criminal justice system, could also warrant ex ante assessments continuously or at regular intervals.

Error Analysis

Algorithms can produce multiple kinds of errors, including careless mistakes, that result from sloppy coding; systematic errors caused by built-in flaws in the algorithm; and random errors stemming from difficult-to-control variations in an algorithm's parameters, the data it uses, or hardware issues.[91] However, errors are to be expected, and their presence should not preclude the use of a particular algorithm, as iteratively improving upon a system by identifying and correcting errors is a standard approach to developing machine learning models.[92]

There are a variety of different error-analysis techniques, including manual review, variance analysis—which involves analyzing discrepancies between actual and planned behavior—and bias analysis.[93] Bias analysis provides quantitative estimates of when, where, and why systematic errors occur, as well as the scope of these errors.[94] While many discussions about the potential harms of algorithms focus on their capacity to replicate human biases, such as racial or gender bias, bias in a statistical context is defined as the propensity to misestimate the value of a particular parameter.[95] This can, of course, manifest as racial or gender bias, but only when race and gender, or proxies for these factors, are involved. For example, an algorithm that generates weather forecasts could be biased toward predicting it will rain on a particularly day, but be exceedingly unlikely to be influenced by racial bias. It is important to bear in mind that algorithmic

bias is a quantitative problem that can cause a wide variety of undesirable outcomes, including but not limited to cognitive biases, such as racial bias.

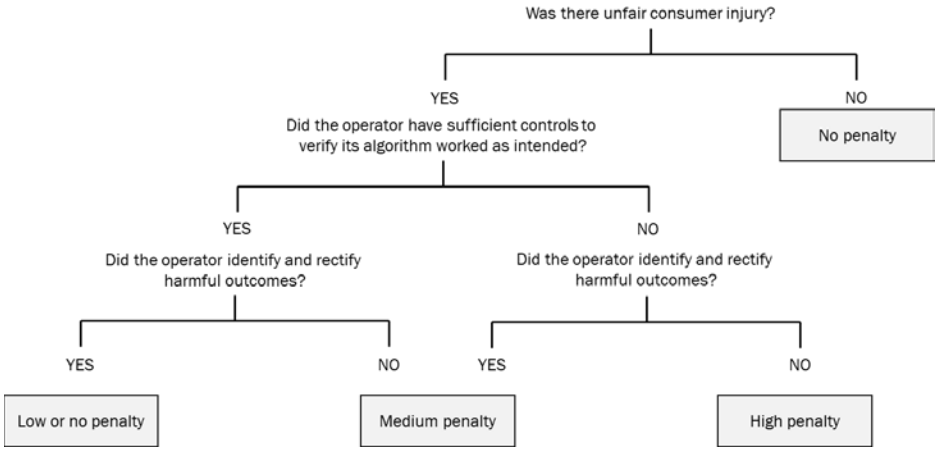
IMPLEMENTING ALGORITHMIC ACCOUNTABILITY

As shown in figure 1, using an algorithmic accountability framework, regulators can take a straightforward approach to evaluating and punishing operators whose algorithms violate existing laws or regulations and produce significant harms worthy of regulatory scrutiny.[96] Importantly, this standard is open to interpretation and will change over time as market forces, social norms, new technologies, and other factors shape the use of algorithmic decision-making.

When operators violate existing laws or regulations using an algorithm, regulators should first examine whether and how effectively the operator can demonstrate they had controls to ensure the algorithm was acting as intended. That operator could be subject to higher levels of punishment if a significant harm occurred and no such controls were present, or if the operator was careless or superficial in their approach to meeting this standard. If these controls were thorough and implemented appropriately, a regulator could likely determine the operator was not acting with negligence or with intent to harm.

At this point, regulators could conduct a similar analysis of whether and how effectively the operator could identify and rectify harmful outcomes. If the operator fails to meet this standard, then a regulator could conclude the operator was irresponsible in their efforts to minimize the potential harms of their algorithm and again be subject to more punishment.

Figure 1: How regulators should use algorithmic accountability



It is important for operators and regulators alike to recognize that although algorithms can produce many kinds of undesirable outcomes, only a certain category of these generate significant harm that warrants regulatory scrutiny. For example, average users of dating apps, social

media sites, and online shopping are unlikely to experience significant harms, even when those online services use woefully flawed algorithms. If algorithms routinely set up users on bad dates, suggest boring links, or recommend unwanted products, users will simply stop using them. Market forces can play a key role in encouraging adherence to algorithmic accountability and would mitigate these harms far more effectively than regulatory intervention. For example, if a company learned the software they purchased to screen job applicants exhibited racial bias, and then shared this information, no other company would buy that software—as not only would the software be provably inaccurate, but sensible companies would fear the public backlash that would result from deploying software with a known racial bias.

In select cases where market forces are muted and significant harm is possible, it may be appropriate for policymakers to dictate specific requirements for algorithmic accountability. This is particularly relevant in the criminal justice system. Caleb Watney, a technology policy fellow at the R Street Institute, argues that because the concept of transparency is central to the goals of the justice system, as indicated by countless court precedents and statutory obligations, such as the Freedom of Information Act and other "sunshine" laws, it would be appropriate to mandate all algorithms that influence judicial decision-making be open-source.[97] Though this transparency may not shed much light on how more-advanced machine learning systems work, there is likely a compelling public interest in ensuring these algorithms are nonetheless exposed to the highest degree of scrutiny possible. Similarly, it would likely be appropriate for policymakers to mandate that public agencies conduct thorough impact assessments for algorithms they intend to use in decisions with high social or economic consequences, such as the administration of entitlement programs.[98] However, any such rules should be narrow and targeted to identifiable harms that algorithmic decision-making could cause in a specific context.

In certain contexts, industry self-regulation could be an adequate means of governance if market forces are diminished or regulation does not apply. For example, while the application of an AI system in health care would typically have to earn regulatory approval, an AI system involved in producing medical research would not necessarily have to do so. Should this system be flawed, doctors applying this erroneous research when treating patients could cause considerable harm. However, the findings of such a system would be subject to peer review just like traditional medical research, thus serving as an effective check against this kind of harm.

By evaluating operators based on this framework, regulators should be able to determine whether operators are negligent in their efforts to prevent harm from occurring. Should a regulator find an operator fails to

meet this standard and causes significant harm, the highest degree of punitive sanctions appropriate for the harm would be warranted. However, it is entirely possible that an algorithmic decision could still cause harm, even with an operator clearly and thoroughly acting in good faith and doing everything that could reasonably be expected of them to prevent this harm from occurring. In such cases, regulators should base their determination of whether to sanction an operator by weighing the harms the algorithm caused against the benefits it generated. Regulators have already condoned this approach to evaluating harms, and it should be easy to apply to harm caused by algorithms. For example, the FTC's 1980 Policy Statement on Unfairness states that the agency would not pursue enforcement action against an unfair business practice if the benefits of the practice—to consumers or to competition—outweighed its harms.[99] This cost-benefit analysis is crucial for ensuring organizations have the ability to develop and use algorithms in innovative ways. If regulators were to presume any harms an algorithm caused—regardless of whether the operator acted responsibly and in good faith—trumped all benefits to society or the economy the algorithm generated and warranted severe sanctions, then operators would simply stop using algorithms.

Importantly, operators should recognize that achieving algorithmic accountability sometimes requires working to identify and rectify harmful outcomes, even if those outcomes are legal. For example, after a self-driving Uber fatally struck a jaywalking pedestrian in Arizona in March 2018, questions arose about whether Uber was legally at fault for the crash.[100] Regardless of whether regulators determined Uber's vehicle conformed to all traffic and safety laws, it is clear this was still an undesirable outcome of algorithmic decision-making. Operators of algorithms in these kinds of high-risk situations should be careful to ensure they can provide their systems with feedback, recognizing that lawful deaths, even those that do not result in legal repercussions for the operator, are still bad and should be avoided. For example, if following an accident, a regulator were to assess whether an operator of autonomous vehicles met the standard of algorithmic accountability, and found that the operator conducted significant testing to ensure its cars were performing as intended and was responsibly identifying and minimizing harmful outcomes, that operator may not deserve sanctions. However, this outcome is still one societies clearly want to avoid, as demonstrated by Vision Zero, an international initiative to eliminate all traffic fatalities and injuries beyond what the law passively tolerates.[101] To deal with these kinds of situations in which algorithms could cause harms, but their operators would not necessarily be breaking the law, regulators could adopt a similar approach to the aircraft-accident investigation process: When a plane crashes in the United States, the Federal Aviation Administration and the National Transportation Safety Board typically conduct an investigation, and even when they do not find that any

regulations were broken, they still publish reports detailing recommendations to improve safety and avoid similar accidents in the future.[102] After this review, operators would have an incentive to implement these recommendations, not only to build safer vehicles, but also to demonstrate that they are responsibly identifying and rectifying potentially harmful outcomes, and thus meet the standard of algorithmic accountability.

Enforcing algorithmic accountability in this way would have important benefits. If operators know this framework exists, they can take proactive steps to ensure they embrace algorithmic accountability, such as by modifying existing systems to increase their transparency, or by discontinuing the use of algorithms that fail to meet these standards. Similarly, this would send a market signal to developers about what customers will expect of an algorithmic system, thus encouraging them to provide algorithms with the necessary capabilities or risk losing market share to competitors that do so.

Importantly, pursuing algorithmic accountability also involves applying existing laws that require transparency, explainability, or other considerations to algorithmic decisions—when relevant. For example, the Equal Credit Opportunity Act requires a creditor to provide consumers with an adequate explanation of why their credit application was declined, while the Fair Credit Reporting Act requires a creditor to both provide consumers with their credit report and investigate disputes concerning incorrect information and make corrections as needed.[103] These laws apply regardless of whether a creditor uses an algorithm in these processes. If a creditor does decide to use an algorithm but is unable to explain why the algorithm declined a consumer's credit application, then the creditor did not meet the standard of algorithmic accountability and is in clear violation of the law.

GOALS FOR POLICYMAKERS

In application areas already governed by existing laws and regulations, policymakers should encourage adherence to the principle of algorithmic accountability. Importantly, policymakers must recognize that the goal of algorithmic accountability is not to achieve perfect, error-free algorithms, but to minimize risk—just as vehicle safety standards do not require cars to be 100 percent safe, but as safe as reasonably could be expected.

The most important step, of course, is for regulators to formally recognize this framework for algorithmic accountability and integrate it into their oversight. This applies to both domain-specific and consumer-protection regulators, as assessing intent to harm and negligence, such as a car manufacturer using an algorithm to deliberately falsify emissions data on their vehicles, or an airline failing to prevent their ticket-pricing algorithm

from price-gouging during a natural disaster, is just as important in manufacturing as it is in health care and consumer finance.[104] This does not mean policymakers should mandate that all algorithms must meet this standard of algorithmic accountability, as it governs the application of algorithms and how to hold operators accountable for violating the law and causing significant harms.[105]

Policymakers should unequivocally reject blanket mandates for algorithms or the creation of new regulatory bodies focused only on regulating algorithms. There are a variety of other steps policymakers should take to support algorithmic accountability, including:

INCREASE THE TECHNICAL EXPERTISE OF REGULATORS

Understanding algorithmic decision-making is necessary for regulators that wish to apply existing regulatory oversight to operators accountable for their use of algorithms. Regulators should foster relationships with communities of developers, academics, civil society groups, and private-sector organizations invested in algorithmic decision-making to stay abreast of technical developments and concerns about algorithmic harms that could influence how algorithmic accountability is achieved or enforced. Additionally, policymakers should ensure regulators have the resources to hire staff with the necessary technical expertise to scrutinize algorithms.

INVEST IN METHODS FOR ACHIEVING ALGORITHMIC ACCOUNTABILITY

Advances in techniques to make algorithms explainable without sacrificing accuracy would make it considerably easier to achieve algorithmic explainability, especially as operators deploy increasingly advanced and complex AI. Policymakers should invest in R&D efforts to support algorithmic explainability, such as DARPA's XAI initiative. Additionally, policymakers should support research into technical methods that could help achieve algorithmic accountability, such as benchmarking systems that can evaluate an algorithm for demographic bias. Finally, policymakers should work with professional societies to support the development of educational materials that could help operators understand how bias or other undesirable factors might influence their algorithms and provide information about how to implement different controls, such as confidence measures or procedural regularity, to combat this.

ADDRESS SECTOR-SPECIFIC REGULATORY CONCERNS

Policymakers should evaluate how effectively this framework for algorithmic accountability would address potential harms in different sectors and consider implementing domain-specific standards for algorithmic accountability where appropriate. For example, it may be necessary to require the criminal justice system to stipulate in its procurement policies that any algorithms involved in judicial decision-

making must be open-source. Additionally, public-sector agencies using algorithms should be required to take into account the significant social or economic consequences of their decisions, conducting thorough and ongoing impact assessments of these algorithms and potentially disclosing this information to the public.

Relatedly, policymakers should make it clear that existing legal and regulatory frameworks still apply to algorithms, just as they do to humans.

CONCLUSION

Some individuals and organizations have attempted to articulate what algorithmic accountability means—and many make good points. However, these definitions often lack specificity or do not effectively convey how to actually regulate algorithmic decision-making. As the economy, and society, become increasingly reliant on algorithms to make both inconsequential and critical decisions, policymakers should carefully consider the benefits algorithms can generate against the potential for these decisions to go awry and cause harm. Unlike the many calls for policymakers to act and regulate algorithms, this framework for algorithmic accountability provides users of algorithms and regulators alike clear rules that can simultaneously maximize the benefits of algorithmic decision-making and narrowly target and prevent harmful outcomes.

REFERENCES

1. 47 U.S.C § 230.
2. Nick Wallace and Daniel Castro, "The Impact of the EU's New General Data Protection Regulation on AI" (Center for Data Innovation, March 2018), http://www2.datainnovation.org/2018-impact-gdpr-ai.pdf.
3. Jeremy Jun, "Big Data Algorithms Can Discriminate, and It's Not Clear What to Do About It," *The Conversation*, August 13, 2015, http://theconversation.com/big-data-algorithms-can-discriminate-and-its-not-clear-what-to-do-about-it-45849; Ramona Pringle, "When Technology Discriminates: How Algorithmic Bias Can Make an Impact," *CBC*, August 10, 2017, http://www.cbc.ca/news/technology/algorithms-hiring-bias-ramona-pringle-1.4241031.
4. Laurel Eckhouse, "Big Data May Be Reinforcing Racial Bias in the Criminal Justice System," *The Washington Post*, February 10, 2017, https://www.washingtonpost.com/opinions/big-data-may-be-reinforcing-racial-bias-in-the-criminal-justice-system/2017/02/10/d63de518-ee3a-11e6-9973-c5efb7ccfb0d_story.html?utm_term=.e1aa634978c3; Julia Carpenter, "Google's Algorithm Shows Prestigious Job Ads to Men, but Not to Women. Here's Why That Should Worry You." July 6, 2018, https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/?utm_term=.3c32398c2f9f.
5. Daniel Castro and Alan McQuinn, "How and When Regulators Should Intervene," (Information Technology and Innovation Foundation, February 2016), http://www2.itif.org/2015-how-when-regulators-intervene.pdf.
6. Daniel Castro and Joshua New, "The Promise of Artificial Intelligence" (Center for Data Innovation, October 2016), http://www2.datainnovation.org/2016-promise-of-ai.pdf.
7. Will Knight, "The Dark Secret at the Heart of AI," *MIT Technology Review*, April 11, 2017, https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/.
8. Cliff Kuang, "Can A.I. Be Taught to Explain Itself?" *The New York Times*, November 21, 2017, https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html.
9. "Humans May Not Always Grasp Why AIs Act. Don't Panic." *The Economist*, February 15, 2018, https://www.economist.com/news/leaders/21737033-humans-are-inscrutable-too-existing-rules-and-regulations-can-apply-artificial?frsc=dg%7Ce.
10. "Algorithmic Accountability" (World Wide Web Foundation, July 2017), http://webfoundation.org/docs/2017/07/Algorithms_Report_WF.pdf.
11. Clare Garvie and Jonathan Frankle, "Facial-Recognition Software Might Have a Racial Bias Problem," *The Atlantic*, April 7, 2016, https://www.theatlantic.com/technology/archive/2016/04/the-underlying-bias-of-facial-recognition-systems/476991/.
12. Ibid.

13. Editorial Board, "Unequal Sentences for Blacks and Whites," *The New York Times*, December 17, 2016, https://www.nytimes.com/2016/12/17/opinion/sunday/unequal-sentences-for-blacks-and-whites.html.

14. Dylan Matthews, "Making Prison Worse Doesn't Reduce Crime. It Increases It." *The Washington Post*, August 14, 2012, https://www.washingtonpost.com/news/wonk/wp/2012/08/24/making-prison-worse-doesnt-reduce-crime-it-increases-it/?utm_term=.b5e14537f123.

15. Kate Crawford, "Artificial Intelligence's White Guy Problem," *The New York Times*, June 15, 2016, https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html.

16. Robert D. Atkinson, "'It's Going to Kill Us!' and Other Myths About the Future of Artificial Intelligence" (Information Technology and Innovation Foundation, June 2016), http://www2.itif.org/2016-myths-machine-learning.pdf.

17. Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St. Martin's Press, 2018).

18. Ibid.

19. Alyssa Edes and Emma Bowman, "'Automating Inequality': Algorithms in Public Services Often Fail the Most Vulnerable," *NPR*, February 19, 2018, https://www.npr.org/sections/alltechconsidered/2018/02/19/586387119/automating-inequality-algorithms-in-public-services-often-fail-the-most-vulnerab.

20. Julia Angwin, Madeleine Varner, and Ariana Tobin, "Facebook Enabled Advertisers to Reach 'Jew Haters'," *ProPublica*, https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters.

21. Dave Lee, "Facebook Can't Hide Behind Algorithms," *BBC*, September 22, 2017, http://www.bbc.com/news/technology-41358078.

22. Martin Baily, Robert Litan, and Matthew Johnson, "The Origins of the Financial Crisis" (Brookings Institution, November 2008), https://www.brookings.edu/wp-content/uploads/2016/06/11_origins_crisis_baily_litan.pdf.

23. "Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights" (Washington, D.C.: Executive Office of the President, May 2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf; Catelijne Muller, *Artificial Intelligence* (Netherlands: European Economic and Social Committee, 2017), https://www.eesc.europa.eu/our-work/opinions-information-reports/opinions/artificial-intelligence.

24. Lee Rainie and Janna Anderson, "Code-Dependent: Pros and Cons of the Algorithm Age," Pew Research Center, February 8, 2017, http://www.pewinternet.org/2017/02/08/code-dependent-pros-and-cons-of-the-algorithm-age/; Christopher Zara, "FTC Chief Technologist Ashkan Soltani on Algorithmic Transparency and the Fight Against Biased Bots," *International Business Times*, April 9, 2015, http://www.ibtimes.com/ftc-chief-technologist-ashkan-soltani-algorithmic-

transparency-fight-against-biased-1876177; "Office of Technology Research and Investigation," Federal Trade Commission, Accessed March 8, 2018, https://www.ftc.gov/about-ftc/bureaus-offices/bureau-consumer-protection/office-technology-research-investigation.

25. Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York: Crown, 2016), ORCAA, Accessed May 8, 2018, http://www.oneilrisk.com/#services-section.

26. Ibid.

27. "Comments of the Electronic Privacy Information Center to the Office of Science and Technology Policy," EPIC, April 4, 2014, https://epic.org/privacy/big-data/EPIC-OSTP-Big-Data.pdf.

28. Lee Rainie and Janna Anderson, "Code-Dependent: Pros and Cons of the Algorithm Age," Pew Research Center, February 8, 2017, http://www.pewinternet.org/2017/02/08/code-dependent-pros-and-cons-of-the-algorithm-age/.

29. Matt Burgess, "Holding AI to Account: Will Algorithms Ever Be Free from Bias If They're Created by Humans?" *Wired*, January 11, 2016, http://www.wired.co.uk/article/creating-transparent-ai-algorithms-machine-learning.

30. Lee Rainie and Janna Anderson, "Code-Dependent: Pros and Cons of the Algorithm Age," Pew Research Center, February 8, 2017, http://www.pewinternet.org/2017/02/08/code-dependent-pros-and-cons-of-the-algorithm-age/.

31. Ibid.

32. Nick Wallace and Daniel Castro, "The Impact of the EU's New General Data Protection Regulation on AI" (Center for Data Innovation, March 2018), http://www2.datainnovation.org/2018-impact-gdpr-ai.pdf.

33. "Humans May Not Always Grasp Why AIs Act. Don't Panic." *The Economist*, February 15, 2018, https://www.economist.com/news/leaders/21737033-humans-are-inscrutable-too-existing-rules-and-regulations-can-apply-artificial?frsc=dg%7Ce.

34. "Algorithmic Transparency: End Secret Profiling," EPIC, Accessed May 8, 2018, https://www.epic.org/algorithmic-transparency/.

35. Cornell Belcher and Dee Brown, "Hailing While Black—Navigating the Discriminatory Landscape of Transportation" (key findings from the hailing while black survey of Chicago voters, Brilliant Corners, February 12, 2015), http://www.brilliant-corners.com/post/hailing-while-black; David R. Francis, "Employers' Replies to Racial Names" (The National Bureau of Economic Research, accessed May 16, 2016), http://www.nber.org/digest/sep03/w9873.html.

36. Nick Wallace, "EU's Right to Explanation: A Harmful Restriction on Artificial Intelligence," *TechZone360*, January 25, 2017, http://www.techzone360.com/topics/techzone/articles/2017/01/25/429101-eus-right-explanation-harmful-restriction-artificial-intelligence.htm#.

37. Ibid.

38. Ibid.

39. "Report: the War on Marijuana in Black and White," ACLU, Accessed May 8, 2018, https://www.aclu.org/report/report-war-marijuana-black-and-white?redirect=criminal-law-reform/war-marijuana-black-and-white.

40. Christian Sandvig et al., "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms" (paper presented to "Data and Discrimination: Converting Critical Concerns into Productive Inquiry," a preconference at the 64th Annual Meeting of the International Communication Association, Seattle, Washington, May 22, 2014).

41. Will Knight, "The Dark Secret at the Heart of AI," *MIT Technology Review*, April 11, 2017, https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/.

42. Ibid.

43. Ibid.

44. Curt Levey and Ryan Hagemann, "Algorithms With Minds of Their Own," *The Wall Street Journal*, November 12, 2017, https://www.wsj.com/articles/algorithms-with-minds-of-their-own-1510521093.

45. Ibid.

46. Patrick Gillespie, "China Broke Hacking Pact Before New Tariff Fight," *Axios*, April 10, 2018, https://www.axios.com/china-broke-hacking-pact-before-new-tariff-tiff-d19f5604-f9ce-458a-a50a-2f906c8f12ab.html.

47. Ibid; Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge, MA: Harvard University Press, 2015), 10.

48. John Faber, "How to Future-Proof Your Search Ranking," Chapter Three, April 2, 2018, https://www.chapterthree.com/blog/how-to-future-proof-your-search-ranking.

49. Ibid.

50. Danny Sullivan, "Google Uses RankBrain for Every Search, Impacts Rankings of 'Lots' of Them," *Search Engine Land*, June 23, 2016, https://searchengineland.com/google-loves-rankbrain-uses-for-every-search-252526.

51. Kate Connolly, "Angela Merkel: Internet Search Engines Are 'Distorting Perception,'" *The Guardian*, October 27, 2016, https://www.theguardian.com/world/2016/oct/27/angela-merkel-internet-search-engines-are-distorting-our-perception.

52. Adam Segal, "Germany Wants Greater Algorithmic Transparency to Fight Disinformation, But Its Approach is Half-Baked," Council on Foreign Relations, April 22, 2018, https://www.cfr.org/blog/germany-wants-greater-algorithmic-transparency-fight-disinformation-its-approach-half-baked.

53. Max Kuhn and Kjell Johnson, *Applied Predictive Modeling* (New York: Springer-Verlag New York, 2013) 50.

54. Jason Brownlee, "Model Prediction Accuracy Versus Interpretation in Machine Learning," Machine Learning Mastery, August 1, 2014, https://machinelearningmastery.com/model-prediction-versus-interpretation-in-machine-learning/.

55. "Turing Lecture: Algorithmic Accountability: Professor Ben Shneiderman, University of Maryland," The Alan Turing Institute, May 31, 2017, https://www.youtube.com/watch?v=UWuDgY8aHmU.

56. "Written Evidence Submitted By the Oxford Internet Institute," Oxford Internet Institute, April 2017, http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/science-and-technology-committee/algorithms-in-decisionmaking/written/69003.pdf.

57. Andrew Tutt, "An FDA for Algorithms," Administrative Law Review 69, no. 83 (March 15, 2016), http://dx.doi.org/10.2139/ssrn.27479944.

58. "Elon Musk, National Governors Association, July 15, 2017," WordsmithFL, July 16, 2017, https://www.youtube.com/watch?time_continue=245&v=b3lzEQANdHk; Camila Domonoske, "Elon Musk Warns Governors: Artificial Intelligence Poses 'Existential Risk,'" *NPR*, July 17, 2017, https://www.npr.org/sections/thetwo-way/2017/07/17/537686649/elon-musk-warns-governors-artificial-intelligence-poses-existential-risk.

59. Alan Ohnsman, "Push for Self-Driving Car Rules Overlooks Lack of Federal Expertise in AI Tech," *Forbes*, July 18, 2017, https://www.forbes.com/sites/alanohnsman/2017/07/19/push-for-self-driving-car-rules-overlooks-lack-of-federal-expertise-in-ai-tech/#2fd44c7dcbf3.

60. Neil Chilson, "How the FTC Keeps up on Technology," U.S. Federal Trade Commission, January 4, 2018, https://www.ftc.gov/news-events/blogs/techftc/2018/01/how-ftc-keeps-technology.

61. "Strategic Human Capital Management," U.S. Government Accountability Office, 2017, https://www.gao.gov/highrisk/strategic_human_management/why_did_study.

62. "Artificial Intelligence and Life in 2030" (Stanford University One Hundred Year Study on Artificial Intelligence, September 2016), https://ai100.stanford.edu/sites/default/files/ai_100_report_0831fnl.pdf.

63. "PM's Speech at Davos 2018: 25 January," U.K. Prime Minister's Office, January 25, 2018, https://www.gov.uk/government/speeches/pms-speech-at-davos-2018-25-january.

64. Rowland Manthorpe, "May's Davos Speech Exposed Emptiness in the UK's AI Strategy," *Wired*, January 28, 2018, http://www.wired.co.uk/article/theresa-may-davos-artificial-intelligence-centre-for-data-ethics-and-innovation.

65. European Group on Ethics in Science and New Technologies, "Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems" (Luxembourg: European Commission Directorate-General for Research and Innovation, 2018), http://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf.

66. Ibid.

67. "Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights" (Washington, D.C.: Executive Office of the President, May 2016),

https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

68. "Keynote Remarks of Commissioner Terrell McSweeny," U.S. Federal Trade Commission, September 10, 2015, https://www.ftc.gov/system/files/documents/public_statements/800981/150909googletechroundtable.pdf.

69. Daniel Castro, "How Congress Can Fix 'Internet of Things' Security," *The Hill*, October 18, 2016, http://thehill.com/blogs/pundits-blog/technology/303302-how-congress-can-fix-internet-of-things-security.

70. Daniel Castro and Alan McQuinn, "The Privacy Panic Cycle: A Guide to Public Fears About New Technologies" (Information Technology and Innovation Foundation, September 2015), http://www2.itif.org/2015-privacy-panic.pdf.

71. Ibid; John Neary, "The Big Snoop: Electronic Snooping—Insidious Invasions of Privacy," *Life Magazine*, May 20, 1966, http://www.bugsweeps.com/info/life_article.html.

72. Tristan Greene, "U.S. Government Is Clueless About AI and Shouldn't Be Allowed to Regulate It," *The Next Web*, October 24, 2017, https://thenextweb.com/artificial-intelligence/2017/10/24/us-government-is-clueless-about-ai-and-shouldnt-be-allowed-to-regulate-it/.

73. Daniel Saraga, "Opinion: Should Algorithms Be Regulated?," Phys.org, January 3, 2017, https://phys.org/news/2017-01-opinion-algorithms.html.

74. Simon Constable, "Why We Should Not Regulate the Tech Industry," *Forbes*, March 26, 2018, https://www.forbes.com/sites/simonconstable/2018/03/26/no-we-really-dont-need-government-regulation-of-the-tech-industry/#2e7ad53deb8d.

75. Christopher Zara, "FTC Chief Technologist Ashkan Soltani on Algorithmic Transparency and the Fight Against Biased Bots," *International Business Times*, April 9, 2015, http://www.ibtimes.com/ftc-chief-technologist-ashkan-soltani-algorithmic-transparency-fight-against-biased-1876177.

76. "Algorithmic Accountability" (World Wide Web Foundation, July 2017), http://webfoundation.org/docs/2017/07/Algorithms_Report_WF.pdf.

77. Nicholas Diakopoulos et al., "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms," FAT/ML, Accessed May 9, 2018, https://www.fatml.org/resources/principles-for-accountable-algorithms.

78. "Algorithmic Accountability" (World Wide Web Foundation, July 2017), http://webfoundation.org/docs/2017/07/Algorithms_Report_WF.pdf.

79. Ibid.

80. Mike Ananny and Kate Crawford, "Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability," New Media & Society 20, no. 3 (December 13, 2016), http://journals.sagepub.com/doi/abs/10.1177/1461444816676645.

81. Caleb Watney, "When it Comes to Criminal Justice AI, We Need Transparency and Accountability," R Street Institute, December 1, 2017,

http://www.rstreet.org/2017/12/01/when-it-comes-to-criminal-justice-ai-we-need-transparency-and-accountability/.

82. David Funning, “Explainable Artificial Intelligence (XAI),” U.S. Defense Advanced Research Projects Agency, Accessed May 9, 2018, https://www.darpa.mil/program/explainable-artificial-intelligence.

83. Cliff Kuang, “Can A.I. Be Taught to Explain Itself?” *The New York Times*, November 21, 2017, https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html.

84. Ibid; Xiaoyan Hu and Phillippos Mordohai, “A Quantitative Evaluation of Confidence Measures for Stereo Vision,” IEEE Transactions on Pattern Analysis and Machine Intelligence 34, no. 11 (November 2012), https://www.cs.stevens.edu/~mordohai/public/Hu_EvalutionOfConfidence_PAMI12.pdf.

85. Joshua Kroll, “Accountable Algorithms (A Provocation),” London School of Economics and Political Science Media Policy Project Blog, February 10, 2016, http://blogs.lse.ac.uk/mediapolicyproject/2016/02/10/accountable-algorithms-a-provocation/; Ibid.

86. Ibid.

87. Organization for Economic Cooperation and Development (OECD), “What Is Impact Assessment?” (OECD, Accessed May 9, 2018), https://www.oecd.org/sti/inno/What-is-impact-assessment-OECDImpact.pdf.

88. AI Now Institute, “Algorithmic Impact Assessments: Toward Accountable Automation in Public Agencies,” Medium, February 21, 2018, https://medium.com/@AINowInstitute/algorithmic-impact-assessments-toward-accountable-automation-in-public-agencies-bd9856e6fdde.

89. Ibid.

90. Travis Korte, “Disparate Impact Analysis is Key to Ensuring Fairness in the Age of the Algorithm,” Center for Data Innovation, January 20, 2015, http://www.datainnovation.org/2015/01/disparate-impact-analysis-is-key-to-ensuring-fairness-in-the-age-of-the-algorithm/.

91. L.C. Nitsche, “Data and Error Analysis,” University of Illinois at Chicago, Accessed May 9, 2018, http://lcn.people.uic.edu/classes/che205s17/docs/che205s17_reading_12a.pdf.

92. Kritika Jalan, “Error Analysis to Your Rescue!” Towards Data Science, January 19, 2018, https://towardsdatascience.com/error-analysis-to-your-rescue-773b401380ef.

93. Frank Wolfs, “Error Analysis,” University of Rochester, Accessed May 9, 2018, http://teacher.nsrl.rochester.edu/phy_labs/AppendixB/AppendixB.html.

94. Lash TL et al., “Good Practices for Quantitative Bias Analysis,” International Journal of Epidemiology 43, no. 6 (December 2014), https://www.ncbi.nlm.nih.gov/pubmed/25080530.

95. “Bias,” Statistics Dictionary, Accessed May 9, 2018, http://stattrek.com/statistics/dictionary.aspx?definition=bias.

96. Joshua New, "When It Comes to Regulating Data, the FTC Has an Economics Problem," Center for Data Innovation, February 15, 2016, http://www.datainnovation.org/2016/02/when-it-comes-to-regulating-data-the-ftc-has-an-economics-problem/; Michael Pertschuk et al., "FTC Policy on Unfairness," U.S. Federal Trade Commission, December 17, 1980, https://www.ftc.gov/public-statements/1980/12/ftc-policy-statement-unfairness.

97. Caleb Watney, "When it Comes to Criminal Justice AI, We Need Transparency and Accountability," R Street Institute, December 1, 2017, http://www.rstreet.org/2017/12/01/when-it-comes-to-criminal-justice-ai-we-need-transparency-and-accountability/.

98. Ibid.

99. Joshua New, "When It Comes to Regulating Data, the FTC Has an Economics Problem," Center for Data Innovation, February 15, 2016, http://www.datainnovation.org/2016/02/when-it-comes-to-regulating-data-the-ftc-has-an-economics-problem/; Michael Pertschuk et al., "FTC Policy on Unfairness," U.S. Federal Trade Commission, December 17, 1980, https://www.ftc.gov/public-statements/1980/12/ftc-policy-statement-unfairness.

100. Timothy Lee, "Video Suggests Huge Problems With Uber's Driverless Car Program," *Ars Technica*, March 3, 2018, https://arstechnica.com/cars/2018/03/video-suggests-huge-problems-with-ubers-driverless-car-program/.

101. "What is the Vision Zero Network?" Vision Zero Network, Accessed May 15, 2018, https://visionzeronetwork.org/about/vision-zero-network/.

102. Sarina Houston, "Inside the Aircraft Accident Investigation Process," The Balance Careers, June 26, 2017, https://www.thebalancecareers.com/inside-the-aircraft-accident-investigation-process-282566.

103. "Rights if Denied Credit," LegalMatch, Accessed May 9, 2018, https://www.legalmatch.com/law-library/article/rights-if-denied-credit.html?intakeredesigned=1; "Disputing Errors on Credit Reports," U.S. Federal Trade Commission, Accessed May 9, 2018, https://www.consumer.ftc.gov/articles/0151-disputing-errors-credit-reports.

104. Russell Hotten, "Volkswagen: The Scandal Explained," *BBC News*, December 10, 2015, http://www.bbc.com/news/business-34324772; Hugo Martin, "Examples of 'Price Gouging' During Hurricane Irma Were Aberrations, Airlines Say," *Los Angeles Times*, September 14, 2017, http://www.latimes.com/business/la-fi-travel-briefcase-price-gouging-20170914-story.html.

105. Daniel Castro and Alan McQuinn, "How and When Regulators Should Intervene," (Information Technology and Innovation Foundation, February 2016), http://www2.itif.org/2015-how-when-regulators-intervene.pdf.

ABOUT THE AUTHORS

Joshua New is a policy analyst at the Center for Data Innovation. He has a background in government affairs, policy, and communication. New graduated from American University with degrees in C.L.E.G. (communication, legal institutions, economics, and government) and public communication.

Daniel Castro is the director of the Center for Data Innovation and vice president of the Information Technology and Innovation Foundation. He has a B.S. in foreign service from Georgetown University and an M.S. in information security technology and management from Carnegie Mellon University.

ABOUT THE CENTER FOR DATA INNOVATION

The Center for Data Innovation is the leading global think tank studying the intersection of data, technology, and public policy. With staff in Washington, D.C. and Brussels, the center formulates and promotes pragmatic public policies designed to maximize the benefits of data-driven innovation in the public and private sectors. It educates policymakers and the public about the opportunities and challenges associated with data, as well as technology trends such as predictive analytics, open data, cloud computing, and the Internet of Things. The center is a nonprofit, nonpartisan research institute proudly affiliated with the Information Technology and Innovation Foundation.

contact: info@datainnovation.org

datainnovation.org