6-16-2021

Automated Decision Making and Machine Learning: Regulatory Alternatives for Autonomous Settings

Alyssa Heminger
Portland State University

Automated Decision Making and Machine Learning: Regulatory Alternatives for Autonomous Settings

by

Alyssa Heminger

An undergraduate honors thesis submitted in partial fulfillment of the

Requirements for the degree of

Bachelor of Science

in

University Honors

and

Political Science and Community Development

Thesis Advisor

Dr. Christopher Shortell

Portland State University

2021

Table of Contents

Abstract

Given growing investment capital in research and development, accompanied by extensive literature on the subject by researchers in nearly every domain from civil engineering to legal studies, automated decision-support systems (ADM) are likely to see a place in the foreseeable future. Artificial intelligence (AI), as an automated system, can be defined as broad range of computerized tasks designed to replicate human neural networks, store and organize large quantities of information, detect patterns, and make predictions with increasing accuracy and reliability. By itself, artificial intelligence is not quite science-fiction tropes (i.e. an uncontrollable existential threat to humanity) yet not without real-world implications. The fears that come from machines operating autonomously are justified in many ways given their ability to worsen existing inequalities, collapse financial markets (the 2010 "flash crash"), erode trust in societal institutions, and pose threats to physical safety. Still, even when applied in complex social environments, the political and legal mechanisms for dealing with the risks and harms that are likely to arise from artificial intelligence are not obsolete. As this paper seeks to demonstrate, other Information Age technologies have introduced comparable issues. However, the dominant market-based approach to regulation is insufficient in dealing with issues related to artificial intelligence because of the unique risks they pose to civil liberties and human rights. Assuming the government has a role in protecting values and ensuring societal well-being, in this paper, I work toward an alternative regulatory approach that focuses on regulating the commercial side of automated decision-making and machine learning techniques.

I. Introduction

For all its promises, artificial intelligence has a major issue: algorithmic bias. In 1997, Amazon's hiring engine was found to penalize words such as 'women's chess club captain' and downgraded applicants who indicated they attended all women's colleges (Dastin, 2018). Though the algorithm was subsequently abandoned, there is nothing that bars companies from employing similar applicant-tracking systems in the future. Similarly, a ProPublica investigation of "risk assessment" scores used by judges to predict recidivism rates among offenders, using social determinants such as race and class, found that the algorithm was prejudiced against Black people (Angwin, Larson, Kirchner, and Mattu, 2016). In each instance, automated decision-making and machine learning techniques assisted in the decision-making process. Each system was given a massive amount of data to quickly sift through, sort, and recognize patterns that would be otherwise unrecognizable by a human. Still, the emerging pattern shows that something went gravely wrong. Identifying whether it was human error during the data "cleaning" process, an unrepresentative dataset, or the implicit biases of the software developer reflected in the algorithm, getting to the root of the issue proves particularly challenging given artificial intelligence, as a machine learning system, involves a process that is not well-understood even by technical experts.

Since the late 1950s, concurrent to the rise of other 'Information Age' technologies, research and development in artificial networks has seen many 'winters' and 'springs.' Beginning in the early 1970s, the Advanced Research Projects Agency (ARPA) of the United States' Department of Defense awarded contracts to develop a system of distributed networks that shared data in real-time. The project, spearheaded by Eisenhower in response to the Soviet launching of Sputnik, became what is known today as the internet (Gabriel, 2020). The Telecommunications

Act of 1996, subsequently passed by the Clinton Administration, revolutionized communication networks by increasing competition in the broadband market, making the internet cheaper and more widely accessible to users across the nation (Ehrlich, 2014). Today, internet users have the advantage of a global network enabling them to be in multiple 'places' at once. However, not unlike most disruptive technologies, several legal and ethical concerns were raised. For example, how does a nation prosecute a crime that happened across borders? In 1994, several hackers, located in multiple continents, hacked into the cash management system of a major U.S. bank funneling $10 million in their personal bank accounts ("A Byte Out of History, $10 Million Hack", 2014). Issues such as these raised basic questions related to jurisdiction and governance prompting a panoply of scholarship by legal academics during the 1990s. Alongside jurisdiction and governance, a host of other challenges were also raised related to concepts such as privacy, security, and intellectual property. The result were laws on copyright infringement, privacy, and hate speech ("Development of the Internet and the World Wide Web," 1999).

Comparable to the rise of the Internet, artificial intelligence was spearheaded by government-sponsored defense programs and privately-funded research projects at large universities and companies ("Development of the Internet and the World Wide Web," 1999). Research into artificial intelligence began in the 1960s but it was not until the 21st century, when required processing power and availability of datasets and labor became available that investment in artificial intelligence as a serious commercial interest boomed. Total global capital investment in artificial intelligence is estimated to be $69.6 billion and growing (Zhang et al., 2021). Advancements in automated decision-making, machine learning, and predictive analysis have streamlined the way industries and various institutions conduct their internal operations and public affairs, further embedding technology into the minds of technologists, researchers,

politicians, scientists, and the public. More and more, businesses are realizing the supply-side advantages of automated decision making, saving both time and money as daily operational efficiency improves. For example, mortgage lenders use decision-support systems to approve prequalified customers for loans in a matter of minutes (Townson, 2020). Still, the scales by which predictive analysis and machine learning are utilized in everyday operations is contextually dependent on the goals of an organization, availability of necessary infrastructure, data and algorithms, and the pace of technological development (Feldstein, 2019).

Admittedly, this is an extremely simplified version of the decades between the initial founding of the Internet to the development of highly sophisticated machines such as IBM's DeepBlue that defeated the reigning world chess champion in 1997. Like the end of the 1990s and early 2000s, which was marked by concern over individual privacy rights spurred by the internet, present day concerns center around the regulation of automated decision-making and machine learning. Many civil society groups, politicians, government, technologists, and researchers, expressing concern over threats to democracy, have called for regulation of technologies that have greater agency in a world increasingly governed by algorithms. Even famed technologist Elon Musk said so at MIT's 2014 AeroAstro Centennial Symposium: "I am increasingly inclined to think that there should be some regulatory oversight, maybe at the national and international level, just to make sure that we don't do something very foolish." Other prominent public figures, Bill Gates and Stephen Hawking, have made strikingly similar statements. This is due to the unprecedented challenges of automated decision-making and its attendant threat to civil liberties and humans rights.

Automated decision-making and machine learning algorithms have an ability to result in unfair bias if preventative measures are not taken such that a statistical overgeneralization is

made, training datasets are unrepresentative of a subset of the population, the implicit biases of the software developer are reflected in the dataset, the operator is susceptible to 'automation' bias, or even a simple error is made during the data input process (Luciano, et al., 2016). While discrimination in every circumstance is not necessarily 'bad,' the black box issue of machine learning makes it nearly impossible to know who to hold accountable when a claim of unfair discrimination does arise. This is because automated decision-making systems are diffuse. That is, the lines of accountability are increasingly blurred when it becomes difficult to identify whether it was the software developer who did not train the machine on characteristically diverse datasets, the judge placing an inordinate amount of faith in the decisional output, the failure of a state agency to use sound data collection methods, or any combination of the three.

In response to these issues, several researchers (see Black and Murray, 2019) claim machine learning can be juxtaposed within existing Information Age regulatory frameworks, especially those concerning the internet. However, while there is considerable overlap, the underlying concerns guiding internet regulation differ in significant ways from those of the present day in regards to automated decision making. Indeed, the commercialization of the internet created the conditions by which these systems began to flourish. For example, the internet introduced efficiency and rapidity into how we communicate while also cheaply commoditizing personal information, laying the foundation for the development of machine learning. However, the shared principles, norms, and rules that shaped responses (or lack thereof) of internet regulation do not provide the framework needed to respond to the unique challenges of artificial intelligence. This is because the government's regulatory strategy, in its approach to ADM, implicitly operates under the same neoliberalist principles that prioritizes innovation in ADM putting principles of transparency, accountability, fairness, and justice at risk.

This paper primarily focuses on the problems associated with unsupervised learning which refers to the ability of machines to quickly sort through and analyze massive amounts of data to find previously unknown patterns in behavior. Emerging patterns from these systems can be used to derive 'actionable insights,' perhaps the most valued form of data to an organization. To help elucidate ADM's unique issues and situate them within a social science, I will offer an epistemological explanation of artificial intelligence and its subcomponents, automated decision-making and machine learning. A brief overview of the technical process through which issues related to unfair outcomes arise will help readers better understand where biases might become embedded, and how developing specific checks or rules along the way could help prevent discriminatory outcomes. In doing so, I hope to move away from the traditional approach that has mainly focused on developing a set of ethical guidelines, or 'soft' laws, that guide automated decision-making and toward a different regulatory approach. I will argue as to why there are several issues with the how the government has approached problems with similar technologies such as the Internet and that ADM with its own unique set of challenges, cannot be regulated in the same manner.

II. Artificial intelligence: Automated decision-making and machine learning

Automated decision-making, machine learning, and predictive analysis are often used interchangeably in referring to a range of computerized tasks. Still, each differs in their operational function. This paper is primarily concerned with machine learning as a highly technical system, capable of performing a broad range of computerized tasks designed to replicate human neural networks, store and organize large quantities of information, detect patterns, and make predictions with increasing accuracy and reliability. A machine learning

algorithm commands a machine to perform a specific set of mathematically complex functions to generate a specified outcome (outputs) given a variety of factors (inputs). The inputs that a machine uses to derive insights refers to data that a business, government agency, or another entity collected either by purchasing it directly from a third-party vendor or entering into a contractual agreement with the user. This type of data can include names, ethnicity, race, and other fixed characteristics. A second criterion of data arguably generates the most controversy due to the method used for its collection and how it is used. Companies will embed cookies into web browsers and track users across webpages collecting data on the number of clicks or the time a user spends on a webpage. In addition to these two types of data, companies are known to collect geographical location data based on users' IP addresses. The private sector's ability to freely collect, use, and even disseminate users' personal information has arguably been the central focus of the discussion over stricter regulation of the internet. I would now like to raise how automated decision-making further complicates this matter and why developing standards for ADM could help deal with shared concerns over these novel systems.

A machine learning algorithm, trained upon a sufficiently large dataset, will build upon previous knowledge by interpolating data and generating novel patterns and insights that would otherwise be unidentifiable by a human (Van Otterlo, 2013). Individualized profiles, either based on factual data–as well as digitally created profiles, or inferred data–are constructed, inferences, made, and users are targeted with advertisements based on what they are likely to be interested in. However, uses of machine learning algorithms often go beyond advertising. Perhaps the most valued form of data, what are called "actionable insights," can be generated whether they be for a CFO projecting future sales or a doctor evaluating the different outcomes for medical treatment decisions (IBM, 2018). These insights, or predictions, that the machine produced can either

augment or entirely replace humans in the decision-making process. For example, operating an aircraft today is largely dependent on knowing how to operate automated flight control systems (Schaper, 2019). Data is generated using negative feedback loops that tell the system if the plane is flying too high or low helping the pilot make adjustments as needed.

This becomes problematic when the data that algorithms depend on are flawed because the algorithm has the ability to reconceptualize and reontologize the world in ways that may be undesirable (Luciano, et al., 2016). Still, even more disconcerting is that machine learning algorithms create a black box. As algorithmic models increase in mathematical complexity, difficulties in conceptualizing the logic of how a machine reached a certain conclusion arise. The machine, given a new set of data, can generate an algorithm that will autonomously define its own operational parameters, or engage in what is called 'unsupervised learning' (Luciano, et al., 2016). The machine generates a classification model for determining how new inputs will be grouped to enable it to make generalizations that go beyond the training data. The matter is complicated by the fact that it is assumed that the operator does not need to understand the rationale behind the development of decision-making rules (Matthias, 2004: 179). The machine is capable of generating algorithms that are sometimes hundreds or even thousands of lines long making it nearly impossible for computer engineers to understand. Even though there is sufficiently large gap between the design and operation of algorithms in complex machine learning techniques, the problem mainly lies with unrepresentative, outdated, or incomplete data, statistical generalizations, and automation bias. (Luciano, et al., 2016).

Research by Hildebrandt (2013) into machine learning methods where humans assisted in decision-making, found that statistical inferences, or generalizations, based on group characteristics and large datasets has had sufficient use in the application of machine learning

outputs. Spurious correlations in proprietary algorithms where the data indicates causation even where the real-world circumstances might indicate otherwise has, in some instances, been enough to act upon the evidence produced by the algorithm (Luciano et al., 2016). This could be indicative of several issues: inadequate training for the operator on the technicalities of these systems such that they can perpetuate or worsen inequalities such that a machine that is trained on an insufficient number of characteristics will have a difficult time recognizing data that deviates from the dataset it was initially trained on. For example, if Black people are more likely to be arrested for a crime because of historical prejudices and racist policing practices this is likely to be reflected in the machine's output (Angwin, et al., 2016). In this instance, the machine would infer that Black people would be more likely to commit a future crime and it is at the discretion of the operator to determine whether the decision is applicable to the person in question.

If the algorithm is not trained on datasets that are representative of the population this could also be a direct result of the software developer incorporating either intentional or unintentional biases into the dataset. For instance, if a Hispanic male software developer only sees the value in training the algorithm on datasets that are representative of the Hispanic population or males then that becomes problematic when the algorithm is used to make decisions regarding Asian women. However, even improper data collection could lead to an error in the data input process if the analyst responsible for 'cleaning' the data makes a simple error. These are examples meant to demonstrate that algorithmic decision-making is fallible to human error at nearly every stage and that a regulatory body composed of experts to monitor the use of machine learning models are imperative to act ex ante and prevent abuses before they arise.

III. Normative concerns and current policy issues

Technology has an integral role in organizing societies, and, like most things revolutionizing production processes, there are likely to be widespread concomitant societal implications. Aside from those wrought by the internet, the turn of the century was marked by a cascade of literature on issues created by automated decision-making. In *The ethics of algorithms: Mapping the debate*, a useful taxonomy was developed to help identify three normative concerns that arise due to the technical shortcomings of ADM: traceability, unfair outcomes, and transformative effects (Luciano, et al., 2016). In their research, the authors concluded that there is significant consensus concerning automated decision-making as a socio-technical system. That is, the values, ethics, and moral concerns that (though oft disagreed upon) provide the context under which these systems technically operate implicate normative and epistemic questions about their ability to aid in achieving societal goals (Luciano, et al., 2016). Realizing the unique challenges that automated decision-making raises, efforts to develop a set of soft laws from normative concerns such as these have matured in recent years, but the literature does not offer much beyond that. Given that bias in automated decision-making has been major point of discussion amongst civil society groups, researchers, regulators, and even technologists, suggesting that there is considerable interest in greater regulatory oversight, an approach that focuses more heavily on applied, data-driven techniques and that does not simply deal in abstraction is necessary in working toward a more salient regulatory framework.

This is not meant to discount concerted efforts made by researchers to translate the technical shortcomings of ADM to broader philosophical precepts; a philosophy is entirely necessary to developing a regulatory framework that transitions away from how the Internet Age technologies have been regulated to a new set of guidelines tailored to the specific challenges of ADM. I am

simply making a broad assertion in saying that many published works (not to the exclusion of legal scholarship) stress the importance that ADM systems are not infallible to human error and, as indicated by Luciano, et. al's research, themes and patterns within the literature are entirely cognoscible. Therefore, greater strides toward a regulatory agenda that builds on this qualitative research should be the next natural progression in the regulatory agenda. This is also not to say the conversation surrounding ADM should cease altogether, but occur simultaneously to greater discussions aiming toward the realization of a set of actionable governmental goals and objectives and the mechanisms available in helping to realize those goals. However, before discussing possible regulatory approaches the government might take to regulate greater use of ADM, I must first discuss problems with the current literature.

More and more, automated decision-making has raised specific concerns not only related to criminal justice but a variety of cross-sector issues such as antitrust and fair business practices, physical safety, personal and societal relationships, economic and social inequality, and more (Edwards and Veale, 2017). Upon recognizing the need for more stringent guidelines, several proposals have been made that situate automated decision-making within broader regulatory frameworks for the internet. However, there are several reasons why the current approach to regulation is insufficient in dealing with the unique issues that automated decision-making systems raise. One, the internet is regulated using a principles-based approach that prescribes principles or guidelines while leaving implementation to individual firms. Principles-based rules introduce ambiguity and subjectivity into the moral or ethical values and are typically used to define behavioral objectives. Typically, principles-based approaches require that 'regulatees' assume greater responsibility (Decker, 2018). For example, Facebook's "Supreme Court," an innovation of CEO Mark Zuckerberg, is a way for users to appeal decisions concerning the ban

of hateful speech on the company's platform (Kang, 2021). This might be unproblematic if the companies that largely develop and control these systems were not currently facing numerous allegations of human rights violations (Google has previously come under fire for its mishandling of sexual harassment claims and for its firing of four employees that were active in labor organizing [Conger and Wakabayaski, 2019]). Moreover, companies could report compliance with standards and guidelines while not actually following them. In other words, there is no enforcement or procedures in place, such as the government conducting an internal audit, to ensure companies follow set guidelines. Where there are no enforcement mechanisms, problems of companies producing falsified reports can (and do) arise. In 2015, the EPA found that Volkswagen had been cheating emissions tests making it appear that diesel cars were emitting far less pollution than they actually were (Clarke, et al., 2015).

Secondly, softer, principles-based forms of law are insufficient on their own for the issues that automated decision-making systems raise because they are slow to adapt to rapidly changing circumstances. There is a growing gap between technological advancement in machine learning (or artificial intelligence more broadly) and, considering partisan gridlock makes it notoriously difficult to pass legislation, laws that reflect changes in attitudes are slow to change. Additionally, it could also be argued that the sector-specific, principles-based approach to internet regulation such as privacy law have largely failed considering that privacy policies are often long and convoluted and users often do not have a choice in which information they choose to disclose without abdicating their rights to freely navigate the internet (Brownlie, 2020). Additionally, computer software is protected by software licensing laws (any software shipped by a company or programmers prohibits the copying or sharing of that software) and proprietary

software that was developed in-house introduces issues of accountability that make it hard for judges to determine who should be responsible when a claim arises.

Still, I must stress that crafting technical standards for automated decision-making and machine learning has proven exceedingly difficult since the problems of regulating automated decision-making come from balancing the promotion of market-based innovation with regulations on the use and application of these systems. Likewise, regulating automated decision-making with a one-size-fits-all approach has been particularly challenging considering automated decision-making operates in different scales, different contexts, and with different abilities (Scherer, 2016). Perhaps then not surprising, there is inherent difficultly in deriving a shared conceptual understanding of these systems and, as a result, strategies for regulating automated decision-making have been piecemeal. For example, the Food and Drug Administration, in taking a risk-based approach to regulating new software as medical devices, has established its own approval process that technologies intending to treat, diagnose, cure, mitigate, or prevent disease must undergo before they are widely distributed ("Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning Based Software as a Medical Device," 2019). However, because the physical threat to safety that software as medical devices pose makes them easier to regulate, automated decision-making is more complex where bias and discrimination cannot be easily proven given a black box. While an array of regulatory tools are available at the government's discretion, the United States government has approached regulation over the last several decades using two dominant approaches: rules-based approach and principles-based approaches. That being said, I believe a hybrid regulatory approach that is more heavily based prescribing rules with additional emphasis given to outlining principles for the regulation of ADM technologies would be best given the unique issues of ADM and machine

learning techniques. This is because traditional principles-based approaches are insufficient on their own given the technical shortcomings and newfangled ethical concerns of ADM systems. Additionally, strict 'command-and-control' approaches, as I will soon show, have their own drawbacks as well.

This paper has proceeded thus far by heavily emphasizing *why* automated decision-support systems require government intervention and less on *how* the government can fulfill its duty of protecting societal well-being from the harms and abuses of these systems. Based on a thorough assessment of the peer-reviewed articles, media publications, government reports, books, and podcasts, missing is a cross-literature bibliometric analysis that identifies the various approaches the United States government might take in realizing its goals provided a synthesis of available regulatory tools. Seeing that regulation does not just exist in the abstract, but as a set of actionable policy goals, the next step should be working toward a regulatory approach.

IV. Toward a Greater Regulatory Framework

Scholarship and the media alike tend to use the term 'regulation' to intuitively refer to government intervention in liberty and choices through an administrative body adopting legal rules or promulgations. At this point, it becomes necessary to define what is meant by regulation. The definitive legal dictionary, Black Law's Dictionary, defines regulation as 'the act of regulating; a rule or order prescribed for management or government; a regulating principle.' Equally alike, the Oxford English Dictionary defines regulation as 'the action or fact of regulating.' If we assume that the goal of regulation is to intervene where the market fails and that the overall governmental objective is to protect societal well-being, then, in fulfilling its obligation, a definition of regulation must include the tools and instruments available to realize

this goal. Therein, the overarching governmental objective for the purpose of a regulatory scheme is as follows: 'Curtailing the threat that profiling places on individuals and groups based on machine-generated insights including data collected on prior and present behavior, personal and professional interests, geographic location, and social determinants such as race, gender, ethnicity, age, and disability.' Executive Order 12866, which requires federal agencies to engage in a cost-benefit analysis of proposed regulations, more narrowly defines regulation as 'an agency statement of general applicability and future effect, which the agency intends to have an effect of law, that is designed to implement, interpret, or prescribe law or policy or to describe the procedure or practice requirements of an agency.' It is the definition that this paper relies on.

Despite current challenges, multiple national and international governing agencies have set forth frameworks for the development of guidelines, principles, and standards for the many uses of ADM systems given underlying ethical concerns. For example, in 2019, Singapore released its Model AI Governance Framework which focuses on articulating a set internal governance structures and measures, human involvement in AI-augmented decision-making, operations management and stakeholder interaction and communication ("Model Artificial Intelligence Governance Framework," 2019). Seeing as the problems of automated decision-making require both technical competence and principled decision-making, an executive agency whose primary purpose is to regulate automated decision-making, and artificial intelligence more broadly, is integral to the development of regulatory schemas. Agencies have a distinct advantage over legislatures and judges in policymaking for several reasons. For one, agency policymakers tend to be experts in the relevant field rather than broad generalists like judges or legislators (Scherer, 2016). Two, agencies have the ability to quickly respond to any changes in the technology

landscape since they are not stymied by the political process. And, third, administrative agencies can act ex ante whereas judges enact laws ex post (Scherer, 2016).

i. Risk-based assessment and cost-benefit analysis (CBA)

To determine the benefits and drawbacks of the potential risks of automated decision-making and machine learning techniques, a risk-based assessment is necessary. Risk-based assessments are well-established in areas of public health and safety, toxicology, environmental regulation, defense, and novel technologies (Scott, et al., 1999). Mathematical techniques analyze the probabilities and potentially harmful effects of an activity making this approach increasingly attractive. Empirically driven by scientific research, risk-based approaches provide decision-makers with probabilities, hazards, and assessments over a particular area concerning society, the environment, economy, government, or industry. These are subjective issues that require the insight of technical experts such as statisticians, data scientists in addition to social scientists, legal scholars, and others. An administrative agency composed of technical expertise would typically develop a risk matrix.

Executive Order 12866, passed by the Clinton Administration, requires that a cost-benefit analysis be prepared for any 'significant' proposed regulation by executive agencies (Sunstein, 2015). These includes costs and benefits to industry, government, individuals, communities, the environment, and the economy. Following a risk-based assessment, I believe a cost-benefit analysis that could include determining the net benefits, harms, and benefits of a specific policy proposal is the successive step in a regulatory proposal. For example, the cost of training tools and materials for the use of automated decision-making systems could be determined using a cost-benefit analysis. Still, many factors could go into a cost-benefit analysis making it hard to

determine the relevant information. For example, how could the social costs of altogether banning artificial intelligence be monetized? Given that the problems of regulation stem from a diffuse technology, regulating automation as a decision-support system must be done through a multi-sector approach in higher education, employment screening, financial services, and healthcare. Therefore, shared consensus amongst technical experts is required further lending credibility to the idea of an executive agency.

i. Rules-based approach

Traditional 'command-and-control,' rules-based approaches are highly prescriptive. They include explicit rules on what can and cannot be done. For example, a 'right to an explanation' forms the basis of the EU's General Data Privacy Rights (GDPR) principles on privacy rights and data usage. A compelling solution to make automated systems more transparent, explainability requires that a logical explanation is provided where the intentions behind the modeling process, summary statistics and descriptions of the input data used to train the model, information on the model's predictive skill, how inputs are turned into outputs, and how the model was tested, trained, or screened for undesirable properties are made available by request (Edwards and Veale, 2017). Explainability stems from concerns over autonomy in greater individual discretion over what type of information is disclosed to private companies. Approaches such as these require that the regulator take responsibility to develop and enforce rules. Say the government wants to focus on preventing unfair outcomes then it could develop a process for showing that a product, service, or system meets the requirements of a technical standard. Such rules that regulators might formulate might require that machine learning algorithms are trained on a specific number of datasets before being approved for use.

However, as argued in *Slave to the Algorithm? Why A 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For*, explainability only goes so far (Edwards and Veale, 2017). For one, meaningful explanations do not work well for every task, and, secondly, explanations fall short where outliers in the dataset cannot be easily interpreted. Looking at the broader intent behind explainability, even by requiring explainability under privacy law, it is not enough for dealing with the larger issues related to the discussion above for two reasons. One, even without information on personal characteristics such as race, or class, machine learning algorithms (as discussed above) draw statistical inferences based on data inputs. Additionally, transparency in the form of explainability only goes so far as the subject is aware that a machine learning algorithm was used to assist in making a decision. Machine learning is used in many contexts that is largely unknown to the user such as in application screenings where the user does not have the option to contest the decision. Therefore, a rules-based approach to regulation might need to be complemented by a principles-based approach that requires employers to disclose if they have used an algorithm for hiring decisions based on avoiding an outcome. Still, applying one approach over another should be based on a critical assessment of the specific context of regulation. Further research might include an empirical study assessing public opinion on the perceived risks and harms these technologies pose.

Figure 1.1: Cost-benefit analysis of differing regulatory strategies in the context of automated decision-making and machine learning

	Costs	Benefits
Rules-based approach	Best where there are high risks Highly prescriptive	High compliance costs Limits scope of innovation

	Limited flexibility Subjective judgements required to assess what actions are likely to achieve the objective	Highly prescriptive Ensures regulator is ultimately accountability Requires ex ante determination of what conduct is permissible
Principles-based approach	Regulatees can overcomply or undercomply Would turn to goals-based approach where risks are heterogeneous such as where risks are diffuse	More adaptive to fast-moving sectors and market changes Encourages experimentation and alternative approaches to compliance Ensures regulatees are accountable
Hybrid-approach	Can 'soften the edges' associated with each approach Combines positive attributes of each approach Allows regulates flexibility and innovation in compliance Allows regulates to better appreciate and understand the general regulatory goals that are being pursued in a specific area	Regulatory strategy could compound the negative attributes of each approach– with neither approaches being efficient nor optimal May not fully reap benefits of each approach

Decker, Christopher, (2018). *Goals-based and Rules-based Approaches to Regulation*. BEIS Research Paper Number 8, Available at SSRN: https://ssrn.com/abstract=3717739

V. Conclusion

Automated decision-making raises specific concerns where the internet does not since the way data is generated, collected, and used has profound implications for what information is valued and what is knowable. With various social actors with competing interests shaping the way data is used and on what terms, private companies are capable of guiding societal behavior

as new patterns in individual behavior, but also institutions, governments, and other entities, emerge. The primary function of this paper is to help researchers operationalize concepts in the literature so far as they are qualitatively focused. I begin with a broad overview of the technical difficulties associated with automated decision-making and machine learning algorithms as they have the potential to exacerbate inequalities in unprecedented ways. Given that widespread concern over these specific issues has been the central focus of the discussion, I argue that researchers have successfully translated these issues into overarching normative concerns arguing that there is a need for some form of regulation that could guide policy objectives set forth by an executive agency. The governmental objective that I propose, given the ethical concerns of these systems, is as follows: 'Curtailing the threat that profiling places on individuals and groups based on machine-generated insights including data collected on prior and present behavior, personal and professional interests, geographic location, and social determinants such as race, gender, ethnicity, age, and disability.' I then proceed by arguing that gaps in the literature exist in identifying specific mechanisms and tools to use in realizing these greater policy objectives. The successive step in strengthening regulatory schemas around automated decision-making comes in deciding how to regulate. A conceptual framework could help industry leaders, researchers and scientists, policymakers, committed to ensuring the protection of human rights and civil liberties, take actionable steps to prevent harms and abuses by these novel systems.

Works Cited

Allo, Patrick, Daniel Mittelstadt, Brent, Floridi, Luciano, Taddeo, Mariarosaria, and Wachter, Sandra. (2016). *The ethics of algorithms: Mapping the debate.* Big Data & Society. DOI: 10.1177/2053951716679679

Angwin, Julia, Larson, Jeff, Kirchner, Lauren, and Mattu, Surya. (2016). *Machine Bias.* ProPublica. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Black, Julia and Murray, Andrew. (2019). *Regulating AI and Machine Learning: Setting the Regulatory Agenda.* European Journal of Law and Technology, 10 (3). ISSN 2042-115X

Brey, P. and Søraker, J. (2009). 'Philosophy of Computing and Information Technology' Philosophy of Technology and Engineering Sciences. Vol. 14 of the Handbook for Philosophy of Science. (ed. A. Meijers) (gen. ed. D. Gabbay, P. Thagard and J. Woods), Elsevier.

Brownlie, Ella. (2020). *Encoding Inequality: The Case for Greater Regulation of Artificial Intelligence and Automated Decision-Making in New Zealand*. Victoria University of Wellington Law Review. Vol. 51 Issue 1, p1-26. 26p.

Cath, Corrine, Floridi, Luciano, Mittelstadt, Brent, Taddeo, Mariarosaria, and Wachter, Sandra. *Artificial intelligence and the 'Good Society': the US, EU, and UK approach.*

Clarke, Sean, Fidler, Matt, Levett, Cath, Scruton, Paul, and Topham Gwyn. (2015). *The Volkswagen Emissions Scandal Explained.* The Guardian. Retrieved from https://www.theguardian.com/business/ng-interactive/2015/sep/23/volkswagen-emissions-scandal-explained-diesel-cars

Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, Yoav Shoham, Jack Clark, and Raymond Perrault, "The AI Index 2021 Annual Report," AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, March 2021. https://aiindex.stanford.edu/wp-content/uploads/2021/03/2021-AI-Index-Report_Master.pdf

Dastin, Jeffrey. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women.* Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Davenport, Thomas H., and Jeanne G. Harris. "Automated decision making comes of age: after decades of anticipation, the promise of automated decision-making systems is finally becoming a reality in a variety of industries." MIT Sloan Management Review, vol. 46, no. 4, 2005, p. 83+. Gale OneFile: Business,

link.gale.com/apps/doc/A134827660/GPS?u=s1185784&sid=GPS&xid=17ace7f2. Accessed 1 Mar. 2021.

Decker, Christopher, (2018). Goals-based and Rules-based Approaches to Regulation. BEIS Research Paper Number 8, Available at SSRN: https://ssrn.com/abstract=3717739

Edwards, Lilian and Veale, Michael. (2017). *Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For*. 16 Duke Law & Technology Review 18.

Federal Bureau of Investigation. (2014). *A Byte Out of History: $10 Million Hack, 1994-Style.* FBI.gov. Retrieved from https://www.fbi.gov/news/stories/a-byte-out-of-history-10-million-hack

Federal Register. (2019). *Maintaining American Leadership in Artificial Intelligence.* National Archives. Retrieved from https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence. E.O. 13859 of Feb 11, 2019

Ferro, David. *A Little Internet History*. Weber State University Department of Engineering, Applied Science, & Technology. Retrieved from https://www.weber.edu/digitalhistory/internet_history.html

IBM Cloud Education. (2020). *What is Artificial Intelligence (AI)?* IBM Cloud Learn Hub. Retrieved from https://www.ibm.com/cloud/learn/what-is-artificial-intelligence

Kang, Cecilia. (2021). *What is the Facebook Oversight Board?* Retrieved from https://www.nytimes.com/2021/05/05/technology/What-Is-the-Facebook-Oversight-Board.html

Info-communications Media Development Authority. (2019). *Model Artificial Intelligence Governance Framework.* Retrieved from https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf

National Research Council. (1999). *Development of the Internet and the World Wide Web*. Funding a Revolution: Government Support for Computing Research. Washington, DC: The National Academies Press. doi: 10.17226/6323.

Presutti, Carolyn. (2013). *FAA Study Issues Recommendations to Correct Pilot Overreliance on Automation.* VOA. Retrieved from https://www.voanews.com/usa/faa-study-issues-recommendations-correct-pilot-overreliance-automation

Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System. Partnership on AI (PAI). Retrieved from https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/

Saporta, Adriel, Patel, Bakul, and Rajpurkar, Pranav. (2021). *The FDA's Bakul Patel on Regulating AI Medical Devices.* The AI Health Podcast.

Schaper, David. *How Pilots Interact with Automation.* (2019). NPR. Retrieved from https://www.npr.org/2019/12/26/791414982/how-pilots-interact-with-automation

Scherer, Matthew. (2016). *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies.* Harvard Journal of Law & Technology: Volume 29, Number 2.

Scott, Alister & Stirling, Andy & Mabey, Nick & Berkhout, Frans & Williams, Chris & Rose, Chris & Jacobs, Michael & Grove-White, Robin & Scoones, Ian & Leach, Melissa. (1999). *Precautionary approach to risk assessment*. Nature. 402. 348. 10.1038/46413.

Shoemaker, Natalie. (2016). *Japanese AI Writes a Novel, Nearly Wins Literary Award*. Big Think. Retrieved from https://bigthink.com/natalie-shoemaker/a-japanese-ai-wrote-a-novel-almost-wins-literary-award

State v. Loomis. (2016). Harvard Law Review. Retrieved from https://harvardlawreview.org/2017/03/state-v-loomis/

Sunstein, Cass. (2015). *Financial Regulation and Cost-Benefit Analysis.* The Yale Law Journal Forum.

Townson, Sian. (2020). *AI Can Make Bank Loans More Fair*. Harvard Business Review. Retrieved from https://hbr.org/2020/11/ai-can-make-bank-loans-more-fair

The White House. (1994). *Executive Order #12866: Regulatory Planning and Review*. Retreived from https://www.archives.gov/files/federal-register/executive-orders/pdf/12866.pdf