Creating Transparency in Algorithmic Processes

*Bernadette Boscoe**

In recent years, innovations such as self-driving cars and image recognition have brought attention to artificial intelligence (AI). Machine learning (ML) is an application of artificial intelligence, made up of algorithms that analyse data and make predictions. Machine learning is increasingly being used to make critical decisions about individuals, such as whether they should be granted parole or whether they are deserving of a bank loan. These new technologies are unregulated, and their processes are insufficiently transparent. Often the predictive algorithms at the core of these technologies are created by private companies and black-boxed, meaning their internal workings are neither subject to nor even open to external oversight. What is worse, even the engineers designing these algorithms do not fully understand how the processes work. As modern society becomes more compute-oriented and data-driven, with computers being used to make critical decisions affecting individuals, greater transparency in algorithmic processes is more important than ever. Any ethical and just society requires some degree of openness and access in its technologies. While it may not be possible to expect an explanation for the entirety of an algorithmic process, we can employ what I will call checkpoints to advance our understanding at various stages of the machine-learning process. Transparent checkpoints can afford policymakers, sociologists, philosophers, information scientists and other non-computer scientists the opportunity to critically evaluate algorithmic processes in the interest of ethical concerns such as fairness and neutrality. The goal of this article is to explain machine learning concepts in a way that will be useful to policymakers and other practitioners. I will give examples of ways biases can be embedded in, introduced into, and reinforced by the original data, and I will introduce six checkpoints wherein black-boxed algorithms can be made transparent, taking care to show how these checkpoints advance our understanding of ethical concerns in machine learning systems.

I. Introduction

In the digital world, algorithms are everywhere. They are sets of computer instructions that perform tasks, such as suggesting movies we should watch, keeping our money safe in the bank, and warning us of objects in our way as we drive a car in reverse. Developments in technology are often presented in a positive light—consider the emphasis on improved safety with self-driving cars and facial-recognition technologies—but algorithms, embedded in computational systems, are also used to make decisions concerning people's lives. We must therefore consider whether the decisions they make are fair or ethical.[1]

Machine learning is a process commonly used to make predictions or to classify people into groups—classifications that have the potential to do harm when, for example, someone is undeservingly rejected for a loan or pays an undue amount for car insurance. Researchers and policymakers are beginning to question machine learning processes, specifically the ways results are obtained, but they have realised it can be difficult to understand how these sys-

DOI: 10.21552/delphi/2019/1/5

* Bernadette Boscoe is a PhD Candidate at the Department of Information Studies, University of California, Los Angeles. This article was adapted from a presentation delivered at the Herrenhausen Conference: Transparency and Society—Between Promise and Peril (June 2018). An extended version of this article was published earlier this year in Henning Blatt et al (eds), *Jahrbuch für Informationsfreiheit und Informationsrecht 2018* (Lexxion Publisher 2019).

1 J Danaher, 'The Threat of Algocracy: Reality, Resistance and Accommodation,' (2016) 29 Philos Technol 3, 245–268

tems work. Companies tend to black-box their tools, not wanting to share their algorithms, while emphasising intellectual property and patent-protection concerns as justifications for their lack of transparency.[2] Software engineers who create machine learning techniques may have little understanding of how their algorithms reach certain decisions, making checkpoints an appealing option for exploring the inner workings of machine learning processes and the relationships between those processes and the outcomes they produce.

While the deepest inner-workings of algorithms might be difficult for most to comprehend, we are certainly capable of evaluating machine-learning processes as a whole and in the interests of fairness and neutrality. I will not define ethical terms such as *fairness* and *bias* in specific ways; rather, I will present a primer for policymakers, social scientists, and ethicists to better understand the machine learning process. I will identify checkpoints, describe what happens at each stage, and provide a bevy of solutions to promote transparency. The citations in the paper are a mix of scholarly and popular articles, and books, mainly intended for non-domain experts. I have attempted to use the latest sources, highlighting groups and researchers looking at these issues, although the literature cited skews to work done within and regarding the United States.

The technical complexity of these tools and systems creates a barrier to understanding their effects on society. We are told that results such as credit ratings are 'accurate,' but we are not told why—or how. Computer scientists are generally not trained in ethical concerns, such as promoting fairness and avoiding bias, and software engineers tend to have little experience in ethical realms. [3] Using algorithms to solve human problems is not new; what is new is the ever-growing datasets and diminished oversight of algorithms that predict human outcomes. Over time, the switch to data-intensive analysis and more complicated approaches to computational problem-solving have engendered a situation where algorithm designers do not fully understand how their own creations work. Older, simpler models were easier to understand; newer algorithms for artificial intelligence are, by contrast, incredibly complex and driven by massive amounts of data.

Against this backdrop, Section II explains machine learning processes for an audience of policymakers, social scientists, philosophers, and others with related interests. Following this, Section III proposes *checkpoints* as a way of promoting transparency and countering the phenomenon of black-boxing. Section IV offers descriptions of six checkpoints that can be made transparent in machine learning processes, checkpoints that can be used when transparency requires human intervention. Ultimately, I argue that creating transparency to understand how these algorithms make decisions is not necessarily incompatible with black-boxing, as checkpoints can clarify processes without exposing trade secrets. In doing so, I will give examples of how transparency can be invoked and instituted in a variety of means to allow the public ways to examine algorithmic systems and, hopefully, lead to the betterment of society.

II. What is Machine Learning?

Machine learning is not new: the term was coined by Arthur Samuel in 1959, and over the years it has slowly evolved to the complex processes used today. Machine learning processes are made up of algorithms, which are themselves made up of code and data. Algorithms are sets of instructions performing tasks in the computer, but these instructions are not explicit; rather, the computer 'learns' from analysing data with known attributes (supervised machine learning) and then makes predictions of similar data.

In the interest of simplicity, I use the term *algorithm* to mean both its conceptual sense and with regard to its implementation. Thus, referring to an algorithm as 'potentially problematic' can mean its execution has produced a result that is ethically troublesome.

In this paper, I will focus on machine learning, as opposed to artificial intelligence, because machine learning methods are commonly used to predict outcomes for individuals, while also designating individuals into categories, which is particularly prob-

2 F Pasquale, *The Black Box Society: The Secret Algorithms that Control Money and Information*, (First Harvard University Press, 2016)

3 C M Pancake, 'Programmers need ethics when designing the technologies that influence people's lives,' (*The Conversation*, 8 August 2018) <http://theconversation.com/programmers-need-ethics-when-designing-the-technologies-that-influence-peoples-lives-100802> Accessed 16 September 2018

lematic. While facial-recognition programs and other deep learning models are also ethically fraught, I will focus on predictive machine learning of human activity, while conceding that these terms can occasionally be blurred (ie, certain ideas can be considered both ML and AI).

In explaining further how ML works, I introduce a thought experiment. Consider the hypothetical Friendly Car Insurance Company (hereafter: Friendly Insurance). When determining the cost of a plan, insurance companies generally look at the type and amount of coverage desired, as well as personal information about the applicant. Assume Friendly Insurance uses machine learning techniques to evaluate a driver's risk potential and to predict appropriate insurance premiums (ie, the cost of a policy). Assume that scientists at Friendly Insurance use a variety of datasets including personal information, while also employing two types of algorithms: some based on open-source software libraries and some that are proprietary entities beyond public access or understanding.

The personal information dataset contains hundreds of thousands of records of individuals that could be past or present customers. The dataset contains *attributes*, also called data *features*, *input features*, or *variables,* describing each individual.[4,5] Each record lists dozens of features about the individual seeking coverage, such as gender, age, and home address. In practice, a data scientist will often choose features to include in or discard from an analysis—thus the 'art' of data science in this regard.

At this point in the process, the data scientist understands what questions she wants the data to answer, and she is ready to assess how much a potential new customer will need to pay for car insurance. This price is called the *target variable*, *target*, or *outcome*.

The data scientist then takes a random portion of the dataset, perhaps seventy percent of the records, and names this the *training set*, before putting the remaining thirty percent into a *test set*. The data scientist then sends the training set (containing all features and output prices) through the machine learning algorithm so it can analyse the features and detect relationships between those features and their corresponding prices. This is the 'learning' part of the computation.

During this analysis, the algorithm discovers from the past examples which features affect premiums, noting how much they negatively or positively affect price. Age, for example, will affect the cost of insurance if the applicant is young. Features are assigned *weights*, depending on how much they factor into prices, and thus if the algorithm finds that prior accidents greatly affects the price, that factor will be weighted more than another feature, such as the color of the applicant's car. The result of this training is a matrix of weights of the features.

At this point, the scientist runs the now-trained algorithm on the test set, and the computer predicts what the premium price should be, based on what it learned from the training set. An accuracy percentage for the algorithm's predictions is displayed, and the accuracy is known because the real premium prices are exposed to the computer and the costs are compared. At this point, data scientists will often run multiple algorithms in search of the best accuracy rate. Once the data scientist is satisfied with the algorithm's efficacy, she can proceed 'into the wild' and attempt to predict premium prices for potential customers.

III. Checkpoints for Transparency

To demonstrate its efficacy as a tool for prediction, I have purposefully chosen a simple example of how machine learning works. The origins of machine learning stem from statistical-based tasks, creating models for prediction.

An increasing number of these tools are being used to decide outcomes based on people, issues which might seem harmless, such as how much an insurance company can charge for coverage. Yet, problems arise in the use of predictive tools. For example, the black-boxing of an algorithm used to assess the likelihood of recidivism, in other words to determine who deserves parole, raises serious concerns about the fairness of techniques deployed within the criminal justice system. In a similarly discon-

4 Terminology varies from field to field; math, statistics, computer science, and data science often use different terms to describe the same concepts. Notation also varies as well.

5 In a ML process, data instances exist as values of feature variables where each feature such as age, height and weight is a dimension of the problem to be modelled. Emre Bayamlıoğlu, 'Contesting Automated Decisions: A View of Transparency Implications' (2018) 4 EDPL, 439

certing way, when consumers are charged more for auto insurance in poorer neighbourhoods, and when the reasons for their higher premiums are beyond (or denied) explanation, bias becomes normalised.[6] The common trope that machine learning is alchemy that 'just works' does not hold up when citizens demand answers.[7] We must then discover how transparency can enable us to further understand these systems.

The solution I propose below is a series of six checkpoints that will expose the inner workings of the machine learning system. Machine-learning techniques are numerous and complex. Some examples include convolutional neural networks (CNNs), adversarial neural networks (ANNs), decision trees, and clustering techniques. Some processes use a multitude of techniques in iterative ways, leading to complicated paths of calculations as the data wend their way through the system. That said, there are ways data scientists can delineate tasks to be performed that can be documented, explained, and understood by the general public. I call these delineations checkpoints to enable transparency. At each checkpoint, various methods are suggested, including activities conducted by both internal and external audits. Ultimately, dividing up the tasks at breakpoints may reveal problems as well as options for addressing them in a targeted manner. Remedies may take the form of reports, metadata, training, governance, and public involvement, among other possibilities. Our task here is daunting, as so few regulations exist, but I will begin with the most important of all, data collection.

IV. Transparent Checkpoints

1. Data Collection

a. Data, Data, Everywhere

All machine learning starts with data. These data may have been obtained from public records, surveys, private data-collection firms, or a combination of sources, and information may have been scraped, munged, simulated, or modelled. Contrary to popular belief, preliminary data are not 'raw'—not mere bits of information obtained objectively; to the contrary, data have been generated through cultural and societal lenses that profoundly and often invisibly shape their nascent existence.[8]

b. Bias in the Data

Since machine learning draws from statistics, computer science, and mathematics, the vernacular used to describe aspects of datasets can often vary. *Bias*, for example, means different things to a statistician and a social scientist.

One of the first problems in examining a dataset is determining its provenance: where, how, and by whom it was created. A major factor leading to biased results in machine learning is that the original dataset is skewed in some way, because machine learning reifies the original data's characteristics and applies them to new data.[9] Since much data is labeled by hand, skewed data can result from many different situations, such as ways that cause certain groups to face bias.[10] Following AI Now Institute's lead, the word *bias* will mean the following for the purposes of this paper: 'Judgement based on preconceived notions or prejudices',[11] with the understanding that bias is also a statistical term with a much more generic meaning. Thus, readers should be cautious with meanings of terms when reading texts concerning machine learning.

Another example of biased datasets is historical data gathered as a result of known, biased police practices. [12] Machine learning techniques will only reify the information they are fed; this is the 'garbage in, garbage out' data conundrum. Scholars Barocas and Selbst[13] look specifically at problems that can emerge in *data-mining* processes, using framing based on American antidiscrimination law. Tools to scrape data from the web are used to gather massive stores of

6 J Larson and J Angwin, 'How We Examined Racial Discrimination in Auto Insurance…,' (*ProPublica*, 5 April 2017) <https://www.propublica.org/article/minority-neighborhoods-higher-car-insurance-premiums-methodology> accessed 29 November 2017

7 M Hutson. 'AI Researchers Allege that Machine Learning is Alchemy' (*Science*, 3 May 2018) <http://www.sciencemag.org/news/2018/05/ai-researchers-allege-machine-learning-alchemy> accessed 8 May 2018

8 L Gitelman, *'Raw Data' Is an Oxymoron* (The MIT Press, 2013)

9 J Zou and L Schiebinger, 'AI can be Sexist and Racist - It's Time to Make it Fair,' (2018) 559 Nature 5, 7714

10 A Campolo et al, 'AI Now 2017 Report' (AI Now 2017 Symposium, New York, 2017)

11 ibid 14

12 K Kirkpatrick, 'It's Not the Algorithm, it's the Data,' (2017) 60 Commun ACM 2, 21–23

13 S Barocas and A D Selbst, 'Big Data's Disparate Impact,' (2016) SSRN Electron J

information. These methods of obtaining data might be violations of terms-of-service agreements or laws, but once data are collected, they are easy to copy and transmit, creating confusion of ownership along legal lines.[14] It is often the case that certain data have been made available online, although they were never intended to be scraped, collected, or combined with other online data to create new datasets.

Returning to Friendly Insurance, we know what kinds of data customers provide to obtain coverage, such as their name, driver's license number, address, and age. But what other information is collected about the individual? In the United States, the external data that can be collected on individuals varies by state. Some states allow credit scores, occupation, and other data to be used to determine premiums, but in reality, all manner of data could be collected, such as shopping habits, criminal records, Facebook posts, and so on.[15] Insurance companies compete to create accurate models of risk so they can attain competitive prices, allowing them to acquire new data in creative ways. But what does all this mean for the consumer?

Another concern in primary data collection is that the data may already be biased, allowing for disparities that will be reified in predictions. For example, as cited earlier, studies have shown car insurance companies in the United States charge higher rates in minority neighbourhoods.[16] Let's assume Friendly Insurance takes the same approach. This means that, later in the machine learning process, the computer will 'learn' that certain neighbourhoods get charged more for premiums, thereby continuing an unjust practice until it is 'told' otherwise.

c. Addressing the Data Collection Checkpoint

Bringing social science methodologies into a computational space can provide a different context and way of understanding bias in datasets. One approach is to hire people outside of computer science and engineering to develop standards and to audit datasets.[17] *Metadata*, for example, can be attached to datasets—think of them as 'datasheets for datasets'[18] —to encourage transparency and accountability by supplying characteristics of the dataset in a standardised way. This approach resembles that of computer hardware, which is required to have detailed information accompanying each component, information that explains its purpose and recommended usage. Examples of items in the datasheet could be data-collection methods, provenance, maintenance, updates and additions to the dataset, privacy conditions, and legal information.[19] Another proposed solution is the Dataset Nutrition Label, a framework that provides a list of dataset 'ingredients', paralleling the approach of food nutrition labels.[20]

Later, when end-results of algorithms are evaluated, this step is crucial in comparing a dataset's intended purpose with its final results. This step is also essential to efforts to root out prejudice that should not have influenced evaluations in the first place.

2. Data-Cleaning Stage

a. What is Data Cleaning?

The data-cleaning stage is the 'in house' data-manipulation stage. In stage one of data collection, the data scientist running the machine algorithms is assumed not to have collected the data, but rather to have simply obtained it from other sources. The data-cleaning stage is where the data scientist studies the dataset(s) to be used, examines them if they need to be altered, and determines what parts of the dataset will be used. For example, some columns might need to be re-labelled or datasets with different labels for the same concepts will need to be streamlined. Data points with little or no data might

14 J Jerome, 'Ethically Scraping and Accessing Data: Governments Desperately Seeking Data' (*Center for Democracy & Technology*, 3 May 2018) <https://cdt.org/blog/ethically-scraping-and-accessing-data-governments-desperately-seeking-data/> accessed 8 September 2018

15 Lately, some car insurance companies offer the option to put a GPS in your car and surveil you, ostensibly for cheaper rates

16 See (n 7)

17 K Crawford, 'The Hidden Biases in Big Data,' (2013) Harvard Business Review <https://hbr.org/2013/04/the-hidden-biases-in-big-data> accessed 28 October 2017

18 T Gebru et al, 'Datasheets for Datasets' Working Paper <https://arxiv.org/pdf/1803.09010.pdf> accessed 26 March 2019. See also S Charrington. 'Datasheets could be the solution to biased AI' (*VentureBeat,* 2 May 2018) <https://venturebeat.com/2018/05/02/datasheets-could-be-the-solution-to-biased-ai/> accessed 26 March 2019

19 ibid

20 S Holland, et al, 'The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards' Working Paper <https://arxiv.org/abs/1805.03677> accessed 26 March 209. See also 'The Dataset Nutrition Label Project 'avail http://datanutrition.media.mit.edu/ accessed 26 March 2019

be removed, and outliers must be addressed. At this step the training sets and test sets are created.

b. Problems in Data Preparation

Choosing what gets included in the data set for analysis is a highly subjective task. The inclusion or deletions of specific features can thus inject certain norms and views into a dataset, norms and views that could introduce or reaffirm biases. Builders of machine learning systems are predominantly white males from Western societies, and they can often fail to realise how their choices might be unfair to certain groups.

As a category, race is problematic. But simply removing race from consideration is, at the same time, rarely sufficient for avoiding racial bias, because race is highly correlated with other categories, such as zip (postal) codes. To this end, the data scientist may potentially impose her bias into the dataset, either consciously or subconsciously. In the case of Friendly Insurance, at this point the data scientist will decide which features to use in the training and test sets. The data scientist's aim is to minimize the company's exposure while finding the right price point for any given consumer. These choices, along with regulations at the state level, shape the model used for pricing decisions, reflecting the broader manner by which outputs are a function of inputs.

c. Solutions to Make Data Preparation Transparent

This checkpoint, like the previous one, involves preparing data for analysis. A report from the data scientist would be combined with metadata gathered about the dataset during the collection phase, perhaps in the form of the above described datasheet. The goals of the analysis are the most important information to document at this point, along with justifications and explanations for the cleaning, ensuring that expressed interests are consistent with the overall process of evaluation. Creating a more robust paper trail that archives the intent of the analysis while also explaining why decisions were made to curate the data in a certain way can help determine if the algorithm is meeting its objectives.

3. Algorithm Choice

a. Choosing an Algorithm

The data scientist may use a number of machine learning methods to analyse data. A particular algorithm might be employed; or, more commonly, a plethora of algorithms might be chosen, with those yielding the best results adopted and deployed.[21] At this juncture, it may be difficult to choose those algorithms likely to yield the best results. I concede the risk in conflating algorithms and the technologies built with algorithms, but given the current climate of little to no oversight over algorithmic processes, any effort to make these systems and structures more transparent is a move in the right direction. Our social systems are fraught with insufficiently understood and woefully undertested machine-learning systems, so accepting the need for reform is a critical first step.[22]

b. Issues with Choosing Algorithms

In the past, easily understood mathematical models were (and are still) used to determine decisions about people's lives. However, more complex machine learning approaches are increasingly being used to automate these processes, meaning that computers themselves have learned how to make decisions in little-understood ways. Current tools lack functionality to provide feedback to explain their decision-making properties, creating clouds of uncertainty around the tools, no matter how intelligent the devices might seem to be.[23]

Because the tools lack ways to explain why and how they make particular decisions, evaluating an algorithm for fairness is difficult. Fairness can mean different things to different people, both ethically and mathematically. A recent study has shown that an ideal form of fairness may be impossible, due to mutually exclusive mathematical properties.[24] Other

21 See for a description of this messy process and why it is so complicated, Pete Warden, 'The Machine Learning Reproducibility Crisis,' (*Pete Warden's blog*, 19 March 2019) <https://petewarden.com/2018/03/19/the-machine-learning-reproducibility-crisis/> 19 March 2018

22 K Crawford and R Calo, 'There is a blind spot in AI research' (2016) 538 Nat News 7625, 311

23 See (n 3) for a good explanation of layers in AI image recognition.

24 G Pleiss et al, 'On Fairness and Calibration' Working Paper <https://papers.nips.cc/paper/7151-on-fairness-and-calibration.pdf> accessed 26 March 2019. Also, see John Rawls 'Justice as Fairness Political not Metaphysical' (1985) Philosophy and Public Affairs 14, 223–51

researchers have identified the problem that prior work on algorithmic fairness looks at conditions for unfairness; but little work has been done to establish measures for fairness.[25] A canonical example of a discrepancy over the definition of fairness between researchers and a software company is the COMPAS risk-assessment tool that was designed to predict rates of recidivism—that is, the likelihood that convicted criminals will reoffend.[26] When researchers evaluating COMPAS disagreed with the company's results, it became clear that the parties were relying on different definitions of fairness in their evaluation of the program. The company claimed fairness on the basis that the scores meant the same thing regardless of the offender's race. On the other hand, the researchers noted that for people who did not reoffend, blacks were twice as likely as whites to be classified as high or medium risk. In this case, the company did not want to explain how its algorithms were coming up with results, so there was no way to examine the algorithm's behavior for why it made the choices that it did. The company relied on end results only to prove its point.

c. Solutions at this Checkpoint

A better understanding of machine-learning algorithms and how they work will benefit society at large. If, for the sake of argument, we assume the original dataset is ethically appropriate and prepared for use, it would be possible to evaluate the algorithm's choices. At this checkpoint, understanding how the algorithm makes decisions can inform data scientists in their algorithm implementations. Also, comparisons of how different algorithms perform and produce results would be useful in decision-making processes.

Mathematical fairness versus human interpretations of fairness (as it pertains to ethical questions) is a tension at this checkpoint.[27] Statistical methods for assessing fairness should be coupled with sociological efforts at defining acceptable parameters for determining fairness. Thus, at this juncture in the machine-learning process, we must recognise and evaluate those mathematical elements that purport to calculate fairness, with special attention to differing definitions of fairness. The parts of algorithms that deal with these notions should be made available for evaluation.

Some approaches involve developing algorithms resistant to human biases in the data; others involve ways of designing algorithms that have demonstrable ways to elucidate algorithms' behaviors in a clear and straightforward manner. A radical, although plausible, solution is to train algorithms themselves to detect unfair decisions and to then remedy the problems detected.[28] However, there is always a danger in asking an algorithm to fix its own troubles.

4. System Design

a. What is System Design?

After the algorithms have been selected, the designer runs the algorithms on the training and test sets, resulting a model to be tweaked and improved upon. For example, a data scientist can adjust weights to certain features to make them more (or less) important within the algorithmic process.[29] Feature selection is a process where the data scientist decides which features will be prioritised in the algorithm.[30] Another way of stating this is that predictor variables are used to assess the value of the target variable. For example, a target variable could be a credit score. Features are inputs, whereas labels, also known as class labels, are system outputs.

b. System Design Can Be Problematic

The combination of human handiwork and machine-learning algorithms can result in a highly complex system that is little understood. Additionally, a number of developers may be involved in adjusting the code along the way, further muddying the waters. Because parameters can be tweaked at any point, dis-

25 T Speicher et al, 'A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices,' Working Paper <https://people.mpi-sws.org/~tspeicher/papers/inequality_indices.pdf> accessed 26 March 2019

26 To access, (n 7) and click on the link in the article 'Auto Insurance Rates are Based on Cost Drivers, Not Race'.

27 J Dressel and H Farid, 'The Accuracy, Fairness, and Limits of Predicting Recidivism' (2018) 4 Sci Adv 1, 5580

28 See (n 10)

29 To drive this point home, tweaking, testing, running and rerunning algorithms is an iterative process; this checkpoint likely comprises many iterations and algorithms, see ensemble learning techniques for further study.

30 M K Lee and S Baykal, 'Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division' (2017) 1035–1048

criminatory results are sometimes only discovered after the processes have been run.[31] The system's opacity is likely greatest at the design stage, because this is when a company's trade secrets are implicated. Opacity can take several forms, such as deliberately concealed inner workings, systems rendered opaque because only technical experts can understand them, and abstruseness due to differences in perception between humans and computers.[32]

For Friendly Insurance this is the point where the data scientist would tweak the model to improve its accuracy. For example, she might find that credit score is a very strong predictor of price point and therefore weight the score more strongly in the model. Insurance is a competitive business, and the company's data scientist does not want to share these ideas with competitors. At this checkpoint the tweaked model is a highly valued trade secret.

d. Proposed Solutions at Checkpoint System Design Stage

Previous work suggests that the form opacity takes determines the technical and non-technical solutions that could help, and thus a first step is to identify types of opacity occurring at this stage.[33] Corporate secrecy concerns certainly require something more than merely not understanding the technology. At this stage of system design, some scholars suggest the use of auditors to examine the code (which would require them being able to access it) to ensure non-discriminatory practices.[34] Considerations should be made if the algorithm works well with certain groups of people but not others, such as marginalised populations, for example, which is a common complaint in the AI sphere. Other work highlights the need to design algorithms that align with the creator's needs, along with ways of designing systems that flag problems and alert creators to potential issues.

5. Results of Algorithm

a. Algorithm Results Description

In machine learning, training sets are fed first through the algorithm and then through test sets to evaluate the models. At this checkpoint stage, before the algorithm is released, its results are examined. This stage assumes that the data scientist is satisfied with the features, weights, and algorithmic choices, and is now evaluating results. Thus, this is the first stage for looking at outputs in a concrete way.

When using test sets, it is possible to obtain error rates on categorical methods. These error rates are calculated by the proportion of cases where the prediction is incorrect. There are many ways to define error in machine learning. I will demonstrate one example from a classification model, where there are four outcomes and where the condition is defaulting on loan. A *true positive* is a result that finds that the predicted condition of the algorithm is true, meaning the person was predicted to have defaulted on a loan and did so. A *false positive* (Type 1 error) occurs when the predicted condition did not happen, meaning the person was predicted to default but did not. A *false negative* (Type II error) is when the condition is not predicted to happen but did, such as when a person is expected to default on a loan but does not. A *true negative*, the fourth possible outcome, is when the condition was not predicted to happen, and did not happen, meaning, in our context, that a person was not predicted to default and ultimately did not.

b. Problems at the Results Checkpoint?

False positive results can have disastrous effects on people. Those individuals who find themselves cast in this light lack a sense of agency in these circumstances; they may not realise they've been categorised as such and, even if they do discover this information, they often lack the power to do anything about it.[35] While credit scores in the United States, for example, are regulated and accessible to consumers, other types of scoring carried out—by companies amassing and combining data from web

31 Center for Democracy & Technology, 'Digital Decisions' (2017) <https://cdt.org/issue/privacy-data/digital-decisions/> accessed 8 March 2019

32 B D Mittelstadt et al, 'The Ethics of Algorithms: Mapping the Debate' (2016) 3 Big Data Soc 2. Also (n 2) scholars claim opacity in the financial sector has become much worse in recent years—that black-boxing has, in other words, gone too far.

33 J Burrell, 'How the Machine 'thinks': Understanding Opacity in Machine Learning Algorithms' (2016) 3 Big Data Soc 1

34 See (n 2)

35 R Courtland, 'Bias Detectives: The Researchers Striving to Make Algorithms Fair' (2018) (*Nature*, 20 June 2018) <http://www.nature.com/articles/d41586-018-05469-3> accessed 10 September 2018

browsing, purchasing patterns, and other sources of public information about individuals—are often used to categorise people in dubious ways.[36]

Because Friendly Insurance is a corporation, the profit motive steers the company to find any advantage it can, using any data it can gather to present the most favorable bottom line to shareholders. Data scientists at Friendly Insurance will almost certainly rely on credit scores generated by third parties, but these scientists generally won't know whether the scores were determined fairly. They incorporated the information but they didn't have a hand in creating it—which is the entire problem in a nutshell. Information is essential—and pervasive—and interested parties will take advantage of the data they need to advance their operations without contemplating the potential injustices built into the system.

As philosophers Mittelstadt et al[37] write, 'algorithmic processing contrasts with traditional decision-making, where human decision-makers can in principle articulate their rationale when queried, limited only by their desire and capacity to give an explanation, and the questioner's capacity to understand it. The rationale of an algorithm can in contrast be incomprehensible to humans, rendering the legitimacy of decisions difficult to challenge.' Faced with these difficult tasks, auditors of systems are wise to individuate the myriad steps and to evaluate pieces the best they can.

c. Solutions

To improve current systems, we require a deeper understanding of what can go wrong in algorithmic processes. Some problems are caused by design mistakes, some by processes gone awry. Some errors, such as false positive classifications, can even be considered a normal part of a machine learning process. Algorithmic systems can be designed to alert humans to processes that appear to be malfunctioning or toward methods for evaluating problematic false positives in results. Oftentimes an algorithm is merely repeating the patterns it finds in the data—which might be inherently biased against certain groups of people—rather than making 'mistakes.' Machine learning system designers must take great care not to employ systems that simply mirror unjust situations.

One solution is to assign responsibility for certain pieces of the system. In this way, designers would be forced to think more deeply about potential outcomes. Institutions can mitigate risk by justifying methodologies and by preserving testing histories for algorithms. Some scholars even argue that public audits of these systems are necessary and should include not only statistical explanations but also related human factors for the safeguarding of privacy rights.[38] Along with this, individuals subjected to the algorithms should be able to inspect their own data, with possibilities for changing this data. Scores, for example, are a singular output when names and other identifying information are entered into an evaluation tool. Legislation is one way to allow individuals to inspect the data associated with them during these processes.

By their very nature, machine learning algorithms evolve and change. It is difficult to determine which versions of code and data to use in making decisions. Creating transparency may seem to be equivalent to instituting more open documentation, but it actually goes beyond that: greater efforts toward transparency will yield greater attention to and an eventual realisation of fairness. If we find otherwise—if transparency indicates that fairness is not possible within the present system of algorithms and the designers creating them—then a more thorough, all-inclusive solution may be required.

Scholarly researchers are looking at new ways to investigate operations of black-boxed algorithms. One line of inquiry deals with *indirect influence*: that is, removing features from the dataset and examining subsequent accuracy rates. In this way, a feature's influence can be quantified.[39][40] These methods involve statistical frameworks that better handle relationships between features and the target. A deeper understanding of relationships between features enables designers to address issues of bias based on feature selection.

---

36 C O'Neil, 'Weapons of Math Destruction' (*Discover Magazine*, 1 September 2016) <http://discovermagazine.com/2016/oct/weapons-of-math-destruction> accessed 13 September 2018

37 See (n 35)

38 D K Citron and F Pasquale, 'The Scored Society: Due Process For Automated Predictions' (2014) 89 Wash LAW Rev, 33

39 Adler et al, 'Auditing Black-box Models for Indirect Influence' Working Paper <https://arxiv.org/abs/1602.07043> accessed 26 March 2019

40 M Hardt et al, 'Equality of Opportunity in Supervised Learning' Working Paper <https://arxiv.org/abs/1610.02413> accessed 26 March 2019

6. End: Evaluation Checkpoint

a. The Evaluation Checkpoint

At the evaluation checkpoint, the machine learning system is operational and has been 'released into the wild.' The system might be in the form of packaged software or a set of black-boxed tools, and it might be employed at a private company, such as an insurance outfit, or in a public institution, perhaps for police or governmental use. Some tools might be built in-house or marketed and sold to others. In the case of Friendly Insurance, potential customers submit their personal information online and then receive a price quote, a quote that is generated through a machine-learning process.

b. Problems at Evaluation

The deployment of a system called Centrelink in Australia's Department of Human Services, which delivers social security payments, illustrates the dilemma of automation.[41] Disbursement was originally done by hand, with people handling the billing services and customer help, but then operations became automated and disaster ensued. Algorithms were riddled with errors, and people were sent bills for debts they didn't actually owe. Adding to the problem, the officers who previously handled customer service were told to steer clients to the computer interface of the system if they desired to challenge these mistakes. Eventually, complaints mounted and the situation became political, with activist groups getting involved and journalists reporting on the events. Researchers can learn from this fiasco by examining how the public interacts with systems, and data collected before and after automating a system can provide insight on these effects.

Replication of results, if possible, is an important consideration at this checkpoint. If the data used in the process is publicly available, other groups may attempt to obtain the same results, thus yielding improved transparency. But, of course, negotiating these black-boxed tools makes the practice of replicating results difficult.

Understanding the machine-learning tool's potential effects on society is critical at this checkpoint. For example, what are the objectives of the tool maker? Has the maker tested for bias? [42] As more of these tools come into use in society, we must trace and analyse their effects. Sets of metrics can be developed to classify them and evaluate how they are deployed in society. Another consideration is monitoring the effects of systems throughout different communities, as bias most often occurs in the data of marginalised communities. Another important point is to track what types of people built and/or will maintain the tool. Is there a diverse representation of people?

c. Proposed Solutions at Evaluation Checkpoint

One solution that has been proposed to engage the ethically fraught environment of machine learning is the creation of *people's councils*. In the words of McQuillian, 'Setting up [such councils] means countering lack of consent with democratic consensus, replacing opacity with openness and reintroducing the discourse that defines due process.' [43] In the context of ethical machine learning, these councils could take the form of other types of councils in health and civil rights organisations. Similar solutions, such as those observed by scholar Crooks, encourage a town-hall-meeting approach to explore technological and ethical concerns by bringing together users, maintainers, and the community to examine the use of technology as a group endeavor.[44]

V. Conclusions

Machine learning systems have increasingly been adopted in a multitude of ways. Technological advancements in ML and AI have thus far outpaced regulatory activities, raising ethical concerns as tools become more widely entrenched in society. There is, however, a growing body of literature calling for ML makers to purposively design systems with potential pitfalls in mind,[45] and also for ML designers to ac-

41 Henman, 'The computer says 'DEBT': Towards a critical sociology of algorithms and algorithmic governance' (*Zenodo*, September 2017) <https://zenodo.org/record/884117#.WvoYYNMvzuR> accessed 26 March 2019

42 See A Breland, 'How white engineers built racist code – and why it's dangerous for black people' *The Guardian* (London, 4 December 2017)

43 D McQuillan, 'People's Councils for Ethical Machine Learning' (2018) 4 Soc Media Soc 2

44 R N Crooks, 'Times Thirty: Access, Maintenance, and Justice' (2018) Sci Technol Hum Values

45 See (n 38)

knowledge differences between benign uses of ML, such as video recommenders, and potentially pernicious uses, such as in the case of bank loans, police surveillance, healthcare, or parole.[46]

The fields of computer science, engineering, and other technical domains are slowly incorporating ethical considerations into design and education realms. Researchers are also attempting to understand precisely how these algorithms work. [47] As evidenced by the footnotes, more people and organisations are recognising that the lack of transparency in machine-learning systems is a serious problem. One major change to data rights has been the rolling out of GDPR, although it remains to be seen what the law's effects will be. Still, it shows steps in the right direction.

The biggest challenge in making machine-learning systems more transparent is the extreme complexity and technological know-how involved. Deconstructing these processes and creating checkpoints is one way to promote greater understanding of black-boxed systems. Collaborations between policy makers, legal scholars, social scientists, and technology experts may encourage more fairness and accountability within machine learning systems.

The general public should also contribute to this effort, sharing their own experiences with these systems. One set of checkpoints alone will not alleviate the problems caused by black-boxed systems; but a combination of solutions, both social and technical, will pave the way for improvements in these systems. Promising work includes standardised solutions, such as datasheets, regulatory changes, and a general shift in thinking toward prioritising understanding of the effects on society in our increasingly algorithmic world.

---

46 C Lewis, 'TSA Will Stop Searching Black Women's Natural Hair (*Essence.com,* 2016) <http://www.essence.com/2015/03/27/tsa-will-stop-searching-black-womens-natural-hair> accessed 8 June 2016

47 N Wolchover, 'New Theory Cracks Open the Black Box of Deep Neural Networks' (*Wired*, 10 August 2017) <https://www.wired.com/story/new-theory-deep-learning> 8 October 2017