Issues in Information Systems
Volume 23, Issue 2, pp. 75-88, 2022

DOI: https://doi.org/10.48009/2_iis_2022_107

A black box approach to auditing algorithms

Seung C. Lee, *University of Minnesota Duluth, slee@d.umn.edu*

Abstract

AI is believed to be the most disruptive force in technology in the coming decade. The best intentions, however, can yield negative consequences, that is, the serious problems of introducing AI, especially algorithmic decision-making and its overtrust, into business and society—the resulting discriminatory and biased decisions. Quite a few studies on algorithmic decision-making have made strong claims about the causality between algorithms and biased decisions. Multiple protective measures, including the Explainable Artificial Intelligence, have also been enacted against the discriminatory and biased algorithmic decision-making practices. Nevertheless, they are persistent because of algorithmic obscurity, biased training data, the false belief that algorithms are neutral, and the public's perception that explainable and data-driven decisions are often not objective. This paper proposes a black-box approach to auditing algorithms. The approach draws on the counterfactual theories of causation. It aims at identifying obvious and obscure decision factors engendering decisions from multiple counterfactuals for a given factual.

Keywords: algorithm audits, counterfactual, algorithmic transparency, algorithmic accountability, bounded rationality

Introduction

Computing machines offer us the means to store, organize, and process vast amounts of data at speed and efficiency that we simply cannot do by ourselves. We have now entered another stage of human innovation known as artificial intelligence (AI) in which the computing machine can learn, reason, and solve problems better in many ways than we can. If this is gotten right, AI can be at the forefront of the next wave of human advancement. Recently, AI has brought business and society major shifts in performing tasks and making decisions. This does not mean AI is a panacea for all the business and societal problems, but there is great expectation around it. AI is believed to be the most disruptive force in technology in the coming decade (Salvatier et al. 2018). The best intentions, however, can sometimes yield negative consequences, that is, the serious problems of introducing AI, especially algorithmic decision-making and its overtrust, into business and society (Guszcza et al. 2018).

Algorithmic decision-making is widely used in both public and private sectors with real consequences for business and society, whether it be an algorithm for assessing the likelihood of a defendant becoming a recidivist (Angwin et al, 2016), a function for determining the risk of undocumented immigrants to public safety (Kalhan, 2013), a formula for dynamic price optimization (Li et al, 2018), or the algorithm-curated information flow (Bandy and Diakopoulos, 2020; McCombs, 2005; Shoemaker and Vos, 2009). Above and beyond, decisions of many kinds are being made by often embedded, connected, and real-time AI-smarts. At the center of these smarts sit algorithms that perform social sorting, job interviews, credit rating, recommendations, premium determination, risk assessment, to name a few. Algorithms have made decision-making processes look handier and efficient, but also have made the process opaque to public scrutiny because they appear as black boxes to the public (Guszcza et al, 2018; Sandvig et al, 2014). Indeed,

a survey of U.S. adults found that the public is concerned about algorithmic decision-making in various real-life situations (Smith, 2018). In the face of important or expensive errors, discrimination, unfairness, or censorship that can be engendered by the decisions made by algorithms, therefore, it is critical to answer the question of how the algorithmic decisions should be accountable to the public. Recently, growing number of studies on algorithmic decision-making calls for public scrutiny of the practice (Asplund, 2020; Guszcza et al, 2018; Raghavan et al, 2020; Sandvig et al, 2014) to achieve algorithmic fairness (Li et al, 2018; Wang et al, 2020) and algorithmic transparency and accountability (Diakopoulos, 2016; Garfinkel et al, 2017). As a recent development, cities around the world are taking initiatives to set policies for using AI technology in providing services such as policing and traffic management. Some require disclosure when an algorithm is used in making decisions, while others mandate algorithm audits or deploy explainable AI (XAI) systems in place (Snow, 2022).

Algorithms are not impulsive, but that does not mean they are neutral when they are making decisions on housing, news feed, job interviews, or policing. As such, algorithmic decisions have huge impacts on many aspects of daily life, but we simply accept the decisions without contesting algorithmic transparency and accountability (Vijayakumar, 2017). Algorithmic transparency is the principle that the factors or variables that are used by algorithmic decision-making should be visible to the regulator or the affected. As Crawford (2013) put it, "if you are given a score that jeopardizes your ability to get a job, housing or education, you should have the right to see that data, know how it was generated, and be able to correct errors and contest the decision." (Angwin, 2016). One noble approach was proposed by DARPA (2017) to mitigate the problems originated from the algorithms posing as black boxes, which is called the Explainable Artificial Intelligence (XAI). Algorithmic transparency can, however, both increase or decrease trust (Bannister and Connolly, 2011). Similarly, XAI could either increase or decrease trust as well, depending on the normative definability of a problem and the complexity of explainability of its solution. This implies that certain strategies are still needed for improving accountability in XAI (de Bruijn et al, 2022).

Imagine we are trying to understand why an animal behaves the way it does. It is tough to figure out why because we can't communicate with such an animal. Similarly, suppose we are trying to understand the rationale behind the algorithmic decisions but the algorithm cannot currently answer on its own why it makes the decisions it does. One of the approaches to solve these problems can be called the deep approach to auditing in which we to understand everything that happens inside the system by understanding in detail the system's different building blocks and how they learn and how those tasks affect the outcomes. Another approach may be dubbed the black-box approach to auditing in which we treat the system as a black box and try to get answers by understanding its behavior only by manipulating its input and figuring out what's interesting in the outputs. Until recently, majority of audit studies have been conducted on non-algorithmic decision-making practices. With widespread use of AI and algorithmic decision-making, researchers eye on algorithm audits in the context of the black-box approach beyond the traditional audit studies. Acknowledging the huge potential impact of algorithmic decision-making on both business and society, we draw on counterfactual theories of causation and on past research to develop a rational counterfactual framework for algorithm audits. We then illustrate the components of the framework to show how it can identify underlying variables and values that can be used to perform algorithm audits to detect or correct algorithmic biases. We conclude with the limitations of the proposed approach.

Theoretical background

Auditing algorithm can be considered to be examining a confluence of three closely related elements: antecedent, algorithm, and consequent. A decision made by an algorithm can be expressed as a factual statement that consists of two parts: antecedent and consequent. In logic, an antecedent is the first half of a propositional statement and a consequent is the second half of it. In algorithmic decision-making,

consequents are conditional upon antecedents and algorithms, but algorithms are in general kept secret from the researcher. We may argue that if the public sector relies on algorithms to make decisions that affect individuals, groups, or whole society, the algorithms and antecedents used to reach the decisions should be visible and explained to the stakeholders. In other situations, we should find a point where we can balance transparency with protecting business stakes and civil interests (Miller, 2015). Although algorithmic transparency is an important aspect of the audit study, it is not only the issue. The more important problem in the audit study is to ensure that the algorithms are applied in a fair and equitable manner beyond making sure that they themselves are fair and equitable (Vijayakumar, 2017).

Auditing algorithms allow researchers to establish causality between the conflation of antecedents and algorithms and consequents (Gaddis, 2018), although algorithms in general are unknown to researchers (Guszcza et al, 2018; Sandvig et al, 2014). Establishing causality is connected to an alternative way of thinking known as the counterfactual theories of causation. In 1748, when defining causation, David Hume mentioned a counterfactual case: "We may define a cause to be an object, followed by another, and where all the objects, similar to the first, are followed by objects similar to the second. Or in other words, where, if the first object had not been, the second never had existed" (Millican, 2007, p. 56). In philosophy and related fields, the fundamental idea of counterfactual theories of causation is that the meaning of causal claims can be explained in terms of counterfactual conditionals of the form "If A had not occurred, C would not have occurred" (Menzies and Beebee, 2019). For example, a well-known audit study of race in the labor market conducted by Bertrand and Mullainathan (2004) found that the resumes with White-sounding names received 50 percent more callbacks for interviews than the ones with African-American-sounding names. This factual statement can be transformed into a counterfactual (or a counterfactual conditional) as: *If the resumes had not contained White-sounding names, they would not have received 50% more callbacks for interviews than the ones with African-American-sounding names*. However, we can come up with a, if not infinite, number of counterfactuals that correspond to a factual (Marwala and Hurwitz, 2017). Suppose we have a factual: *Rigorous lockdowns were put in place to slow the spread of the coronavirus and consequently they prevented tens of millions of infections and saved millions of lives*. Its counterfactual can be: *If rigorous lockdowns had not been put in place to slow the spread of the coronavirus, tens of millions of people would have been infected and millions of lives would not have been saved* or *If modest lockdowns had not been put in place to slow the spread of the coronavirus, tens of millions of people would have been infected and millions of lives would not have been saved* or *If gentle lockdowns had not been put in place to slow the spread of the coronavirus, tens of millions of people would have been infected and millions of lives would not have been saved*. It is obvious that there are many different ways in which we can formulate counterfactuals for a given factual.

Counterfactuals have been used for decision making and are essentially a process of comparing real and hypothetical situations and using their difference to make decisions (Cantone, 2020). Counterfactual analysis is a powerful framework that can be used to prevent future disasters. For example, there is a factual that the technicians at the Chernobyl Nuclear Power Plant conducted a poorly designed experiment on a reactor with design flaws and, consequently, there were explosions and fires releasing large amounts of radioactive material into the atmosphere that killed tens of people, caused tens of thousands people to have thyroid cancers, and forced hundreds of thousands of people to be evacuated. We can identify conditions that could have led to the prevention of the Chernobyl catastrophe and use the information to prevent future similar accidents. There are a number of ways in which counterfactuals can be formulated using structural equations (Woodward and Hitchcock, 2003). Within a collection of counterfactuals that correspond to a given factual, there can be a number of counterfactuals that maximize the utilities of particular consequences. Such counterfactuals were dubbed by Marwala (2014) rational counterfactuals in line with the theory of rational choice of economics in which individuals make decisions based on their preferences among the available alternatives. However, achieving maximum utility may not be feasible because

rationality of individuals is limited by the available information, the cognitive constraints, and the time limits in making decisions. Therefore, the rational counterfactuals in fact should be understood as bounded ones in line with the theory of bounded rationality (Marwala and Hurwitz, 2017).

Audit studies allow researchers to make strong causal claims (Gaddis, 2018), and the fundamental idea of the counterfactual theory is that the meaning of causal claims can be explained in terms of counterfactual conditionals. In other words, one use of counterfactual conditionals is to define causality (Menzies and Beebee, 2019). This close connection between audit studies and counterfactual conditionals is highly relevant to the context of auditing algorithms. According to Hempel's deductive-nomological (D-N) theory of explanation, explanations have the logical form of two major components (Hempel and Oppenheim, 1948): a sentence describing the phenomenon, termed an *explanandum*, and the group of those sentences that are cited as evidence to account for the phenomenon, termed *explanans*. A couple of conditions must be met to ensure that the explanans successfully explain the explanandum. First, "the explanandum must be a logical consequence of the explanans" and "the sentences constituting the explanans must be true" (Hempel and Oppenheim, 1948, p. 137). That is, the phenomenon to be explained must be logically deducible from the particular circumstances or initial conditions. Second, the explanans must contain at least one proposition that expresses a regularly occurring or inevitable phenomenon, that is, at least one law of nature, and this must essentially be included in the derivation or deduction of explanandum from explanans. Otherwise, the derivation would be invalid without this premise. Many phenomena are explained by the D-N theory. For example, slowing the spread of the coronavirus can be explained by the particular circumstances including maintaining social distance, wearing respiratory masks, avoiding person-to-person interactions for extended periods, and washing hands.

However, generalizations that either conform to the D-N model or are not plausibly deemed laws of nature can also be used to answer a range of what-if-things-had-been-different questions as long as they have the right sort of invariance characteristics (Woodward, 1996; Woodward and Hitchcock, 2003). In an abstract sense, the generalization in the sense of the D-N model not only shows that the explanandum is explained under the given particular circumstances or initial conditions but it can also be used to show how this explanandum would change if the circumstances or conditions were to change in various ways. Stated differently, counterfactual conditionals can be used to show how consequents change in appropriate ways with interventions (e.g., parameterization or transformation) on the antecedent variables. That is, counterfactual conditionals can answer a range of what-if-things-had-been-different questions without citing laws of nature (Marwala, 2014; Woodward, 1996; Woodward and Hitchcock, 2003).

Counterfactual theories of causation

As stated earlier, the basic idea behind the counterfactual theories of causation is that counterfactuals can explain the relationship between cause and effect of certain factuals (Menzies and Beebee, 2019). Counterfactual thinking has brought about difficult semantic, epistemological, and metaphysical questions: a semantic question, as how do we communicate and reason about possibilities which are far from the way things actually are?; an epistemic question, as how can our experience in the real world justify the reasoning about remote possibilities?; a metaphysical question, as do these far-off possibilities exist independently from the real world, or are they hinged on things that actually exist? (Starr, 2019). Nevertheless, counterfactual analyses have become popular since the best-known counterfactual theory of causation by Lewis (1973). Lewis's study is given credit for the best known and most thoroughly elaborated counterfactual theory of causation so far. Lewis succinctly described the underlying idea behind counterfactual analyses of causation as "We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it. Had it been absent, its effects – some of them, at least, and usually all – would have been absent as well" (1986, pp. 160-161).

Lewis employs the semantics of possible and actual worlds for counterfactuals to elaborate truth conditions for counterfactuals in terms of a comparative similarity relation between possible and actual worlds. One possible world is said to be closer to actuality than another if the former resembles the actual world more than the latter does. By means of this comparative similarity relation, the truth condition for the counterfactual "If *C* were (or had been) the case, *E* would be (or have been) the case" is stated as follows: If *C* were (or had been) the case, *E* would be (or have been) the case" is true in the actual world if and only if either (1) there are no possible *C*-worlds; or (2) some *C*-world where *E* holds is closer to the actual world than is any *C*-world where *E* does not hold. In other words, the counterfactual "If *C* were (or had been) the case, *E* would be (or have been) the case" is true if only if it deviates less from actuality to make the antecedent true along with the consequent than to make the antecedent true without the consequent.

With regard to counterfactuals, Lewis defines a notion of causal dependence between possible events, which plays a central role in his theory of causation: Where *c* and *e* are two distinct possible events, *e* causally depends on *c* if and only if, if *c* were to occur *e* would occur; and if *c* were not to occur *e* would not occur. This condition states that whether *e* occurs or not depends on whether *c* occurs or not. Where *c* and *e* are events that actually occur, this truth condition can be simplified following the second formal condition on the comparative similarity relation above. That is, the counterfactual "If *c* were to occur *e* would occur" is automatically true and this implies that a counterfactual with true antecedent and true consequent is itself true. Consequently, the truth condition for causal dependence becomes: Where *c* and *e* are two distinct actual events, *e* causally depends on *c* if and only if, if *c* were not to occur *e* would not occur. This definition of causal dependence is based on three important premises. First, it primarily deals with events where relations exist between them, although it is possible to formulate causal dependence in terms of facts rather than events (Mellor 1995, 2004). Second, the definition requires the causally dependent events to be distinct from each other. That is, the events are not identical, neither is part of the other, and neither implies the other. Third, the right counterfactuals to be used are non-backtracking counterfactuals. For example, suppose that the events *c* and *e* are effects of a common cause *d*. Then, the right counterfactuals make any causal dependence between *c* and *e* void so the inference to the claim that *e* causally depends on *c* is blocked.

Ever since the Lewis's study, extensive exploration of the theory over almost fifty years has called into question about the adequacy of any simple analysis of singular causation in terms of counterfactuals (Elga, 2000; Hall 2004; Paul and Hall, 2013). Consequently, recent years have witnessed the development of an alternative counterfactual approach to causation that employs the structural equations framework (Hitchcock 2001, 2007; Woodward 2003; Woodward and Hitchcock 2003) that is currently the most popular way of analyzing the relationship between causation and counterfactuals (See Hitchcock, 2001).

Counterfactual lake and rational counterfactuals

Aforementioned, there is a counterfactual for every combination of possible values of antecedent variables. Therefore, in general, we can evaluate a counterfactual, say "If it were the case that $X_1,\ldots,X_n$, then …", by replacing the original equation for each variable X_i with a new equation specifying its hypothetical value, while keeping the other equations unchanged; then the values for the remaining variables are calculated to see whether they make the consequent true. This technique of evaluating an equation with a new hypothetical value set by a *surgical intervention* describes the concept of counterfactual dependence between variables as follows:

> A variable Y with its value y *counterfactually depends* on a variable X with its value x in a causal model if and only if there exist values $x' \neq x$ and $y' \neq y$ such that replacing the equation for $X = x$ with $X = x'$ yields $Y = y'$.

This definition implies that there can be as many counterfactuals as the number of possible values of the antecedent variables, which in turn suggests that a factual can be transformed into a collection of counterfactuals that may be called a *counterfactual lake*. For example, suppose we have a factual: *Billy opposed wearing masks and keeping social distance and consequently he contracted the coronavirus*. Its counterfactual can be: *If Billy did not oppose wearing masks and keeping social distance then he would not have contracted the coronavirus* or *If Billy did not oppose wearing masks he would not have contracted the coronavirus* or *If Billy did not oppose keeping social distance he would not have contracted the coronavirus* or *If Billy occasionally opposed wearing masks and keeping social distance then he would not have contracted the coronavirus* or *If Billy opposed wearing masks and keeping social distance once in a while then he would not have contracted the coronavirus*. This narrative clearly shows that there are many different ways in which one can formulate counterfactuals for a given factual.

Within a counterfactual lake that corresponds to a given factual, there can be a number of counterfactuals that maximize particular utilities. Such counterfactuals are called rational counterfactuals in line with the theory of rational choice of economics in which individuals make decisions based on their preferences among the available alternatives. (Marwala, 2014). The notion of rationality here refers to bounded rationality that departs from the assumptions of perfect rationality of homo economicus. The perfect rationality assumes an economic agent who has complete information about the options available to choose from, perfect knowledge of the consequences from choosing those options, and the means to solve an optimization problem that identifies an option which maximizes the agent's personal utility. However, achieving maximum utility in making decisions is bounded by many factors such as available information, cognitive constraints, and time limits (Simon, 1957).

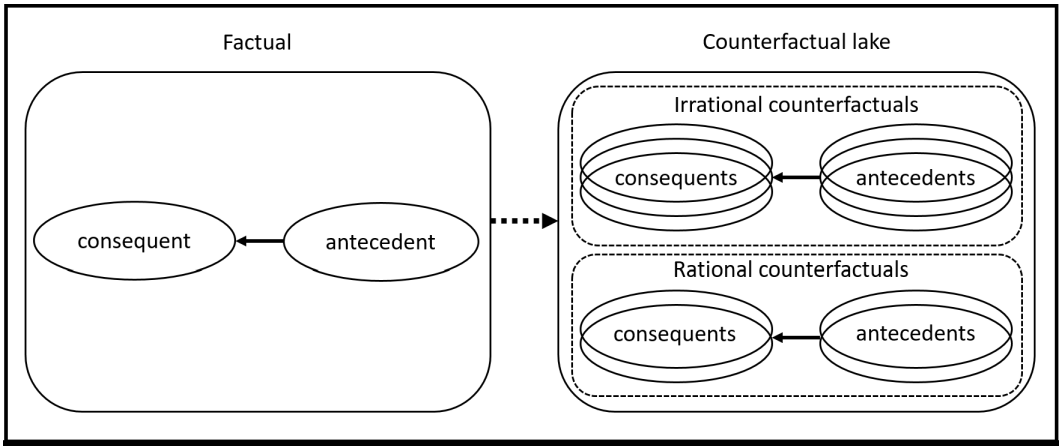

Figure 1. Transformation of a Factual into a Counterfactual Lake of Irrational and Rational Counterfactuals

Figure 1 illustrates a transformation of a factual into a counterfactual lake that contains both irrational and rational counterfactuals. A factual statement is comprised of antecedent and consequent and read from right to left, which is indicated by a leftward arrow. The same holds true for the counterfactuals of the counterfactual lake for the given factual. As shown, the counterfactual lake consists of a set of irrational counterfactuals and another set of rational counterfactuals and the multiplicity of counterfactuals is shown by ellipses.

Counterfactuals and audit study designs

Audit study has long been used for research in various disciplines but there had been no proposed algorithm audit study designs until Sandvig et al (2014) offered five of them that can also be used to examine the normative problems (e.g., race-based discrimination in housing) brought up earlier in this paper. However, there are advantages and drawbacks of each audit study design. The disadvantages are largely related to technical issues such as sampling of participants, validity of design, research ethics and legality. As such, the proposed audit study designs focus on the technical issues rather than how to look into an algorithmic decision as a factual that is comprised of antecedent and consequent. The domains of previous audit studies have largely been race, ethnicity (e.g., Daniel, 1968), and gender (e.g., Levinson, 1975), but recently, they have been expanded into age (e.g., Bendick et al, 1997), criminal record (e.g., Evans, 2016), disability (e.g., Baert, 2014), educational credentials (e.g., Deming et al, 2016), immigrant assimilation or generational status (e.g., Gell-Redman et al, 2017), mental health (e.g., Baert, 2016), military service (e.g., Figinski, 2017), parental status (e.g., Bygren et al, 2017), physical appearance (Ruffle and Shtudiner, 2015), religious affiliation (e.g., Wallace et al, 2014), sexual orientation (e.g., Mishel, 2016), social class (e.g., Rivera and Tilcsik, 2016), courtroom algorithm (Hao and Stray, 2019), and news curation systems (Bandy and Diakopoulos, 2020).

One of the common characteristics of the studies is that they use one or two antecedent variables to make causal claims on the consequent. However, as described before, there can be many number of counterfactuals for a given factual and each counterfactual can be formulated for every combination of possible values of more than one or two antecedent variables. In other words, when applied to algorithm audit, a decision made by an algorithm can be expressed as a factual statement that consists of two parts, antecedent and consequent, and the factual statement can be transformed into a counterfactual lake containing both irrational and rational counterfactuals. This leads to a more general audit study design for auditing algorithms, called a rational counterfactual framework. It can take multiple antecedent variables, either obvious or obscure or both, into consideration. It is also in line with Lewis's semantics of possible and actual worlds (1973) and Hitchcock's counterfactual approach to causation known as the structural equations framework (2001).

A framework for black box approach to auditing algorithms

Algorithmic-decision making system subject to audit

The proposed rational counterfactual framework for auditing algorithms, as illustrated in Figure 2, starts with the following assumptions: (1) an algorithmic decision, including an XAI decision, can be transformed into a factual statement that is comprised of antecedent and consequent; (2) the antecedent of a factual can be elaborated into a collection of variables that can be *intervened*; (3) a number of counterfactuals can be formulated for a given factual by a surgical intervention on each antecedent variable; and (4) a counterfactual lake is comprised of rational and irrational counterfactuals, where the notion of rationality here refers to bounded rationality.

The possible biases and discrimination led or induced by computer algorithms differ from the non-algorithmic counterpart processes in a number of crucial ways (Sandvig et al, 2014). First, algorithms can affect large number of people. Second, algorithms mostly remain as black boxes (Guszcza, 2018). Even if they are disclosed, it does not mean they can be interpreted by reading the code. Even an expert may not be able to predict how the algorithms would behave without testing with some example data and examining the results. Third, algorithms sometimes disproportionately depend on private data as inputs. As a result, the same algorithmic decision may never be made twice. Finally, there is no reason to believe that the

algorithms will act in the best interests of the affected in the absence of regulatory oversight. Thus, any computer algorithms that drive decision making may be audited to ensure they do not exhibit bias. For example, the proposed Consumer Online Privacy Rights Act (COPRA) would force companies to audit the decisions made by any covered AI/ML systems in an effort to mitigate bias and other potentially negative consequences of automated decision-making. The CORPA's requirement would be in line with GDPR's requirement for algorithms to *implement technical and organizational measures* that *prevent, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect* to combat algorithmic discrimination (Goodman, 2016).

Antecedent and consequent of a factual

The term "factual" is defined as the thing that is actual or real. It is concerned with facts that are in general independent of belief. In this paper, a factual means a decision made by an algorithm whether or not it is biased, while a factual statement refers to a factual presented in detail with antecedent and consequent. For example, the factual statement "the resumes with White-sounding names received 50 percent more callbacks for interviews than the ones with African-American-sounding names" contains a fact "50 percent more callbacks for interviews" as a consequent resulted from the antecedent of "the resumes with White-sounding names." Although it is possible to separate a factual from a factual statement in lexicographical sense, the term *factual* is used interchangeably with the phrase *factual statement*. Examining a factual allows us to identify, implicitly or explicitly, antecedent variables, obvious or obscure, that could engender the consequent incorporated in the factual. The primary tool for fighting algorithmic biases is to sanitize data used in automated decision making, that is, to prevent the inclusion of antecedent variables related to protected categories, including race, gender, age, religious affiliation, and sexual orientation. This can be called the *basic requirement* for a decision-making algorithm. In addition, it is also important to not include any other antecedent variables of non-protected categories (or proxy variables for the protected category variables) if they, individually or jointly, have a statistically significant relationship with the protected category variables. This can be referred to as the *extended requirement* for an algorithm. However, both requirements are far from being done for fighting algorithmic biases.

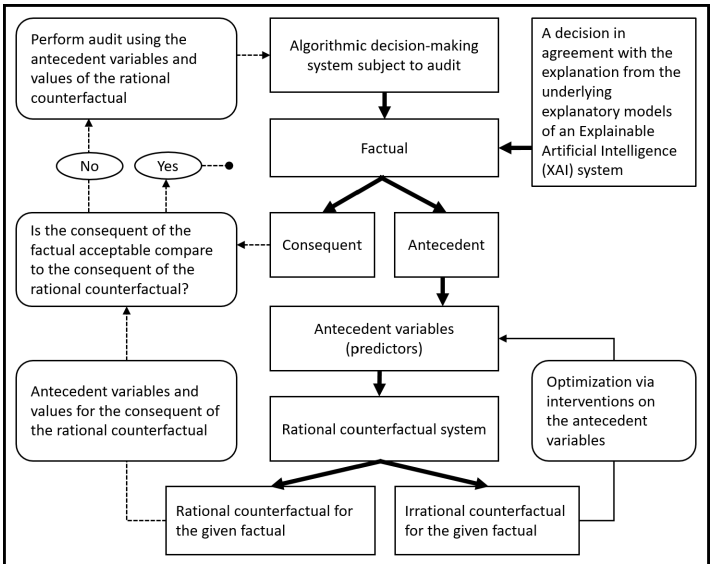

Figure 2. A Rational Counterfactual Framework for Auditing Algorithms

An algorithm would be initially fit on a training dataset. It will, however, acquire a *taste* for discrimination if the relationship between the dependent and independent variables in the training dataset mirrors noticeably discriminatory treatment (Becker, 2010; Custers et al, 2010). For example, if an algorithm is trained on the dataset of past performance ratings that are racially biased and race is explicitly coded, the resulting algorithm will definitely discriminate on the basis of race. In this case, the basic requirement would require dropping the race variable from the dataset. In addition, the basic requirement is not effective in cases of statistical discrimination, where an antecedent variable related to the protected categories is genuinely predictive. Then, encoding the variable in the algorithm would become redundant (Barocas and Selbst, 2016). The extended requirement attempts to remove both explicit and proxy variables for any of the protected category variables. However, removing all proxy variables will likely end up with the loss of useful information for decision-making process (Calders and Verwer, 2010). Furthermore, eliminating all variables that have a statistically significant correlation with a protected category variable from a dataset does not guarantee that remaining variables will not interact with a protected category variable on the aggregate (Dodge, 2003).

Rational counterfactual system

A rational counterfactual system can be any system that performs at least two basic functions: a function to generate counterfactuals, either desired or undesired according to social norms, laws, or economic standards, for a given factual and another function to optimize the undesired counterfactuals within a simulated setting. A rational counterfactual system may be a machine learning system that consists of sophisticated models capable of representing complex, non-linear decision boundaries or simple structural equation models that link the consequent of the model and the antecedent. Such a machine learning system may employ techniques such as neural networks and fuzzy logic with computational method like particle swarm optimization or genetic algorithm (Marwala, 2014).

As shown in Figure 2, a rational counterfactual system generates counterfactuals, each of which is the result of interventions on the antecedent variables and can be either rational or irrational in terms of its consequent. Finding a rational or desired counterfactual may require multiple iterations. For example, one can assume that supposing the social distance among people is 3 feet, then what will be the transmission rate of the coronavirus, and, the model will be able to give a transmission rate say 5%. Then one can imagine another counterfactual, say supposing the social distance among people is 6 feet and the model then gives a transmission rate of 1%. This process can be repeated until a desired transmission rate is achieved with each iteration of a different counterfactual. Once a desired counterfactual is determined, its consequent is compared with the counterpart of the factual from which the desired counterfactual has been generated. If the consequent of the factual is not acceptable compared to the consequent of the desired counterfactual, then the algorithmic decision-making system will be audited using the antecedent variables and values that have led to the rational counterfactual.

Conclusions and limitations

In this paper we discussed the danger of algorithmic decision-making practice that is emanated from the possibility of making discriminatory or biased decisions. We then introduced the social scientific audit study, a methodological tool considered to be the most rigorous way to test for discrimination and biases in many high-impact business and social domains such as employment and housing. After outlining some of the challenges and limitations of traditional audit studies and the existing methods for algorithm audits, we proposed a framework for auditing algorithms as a research tool that is founded on the counterfactual theories of causation. We also discussed the possibility of alleviating if not eliminating algorithmic biases

by achieving the two principles in algorithmic decision-making, namely algorithmic transparency and data sanitization. However, attaining algorithmic transparency is not highly feasible because algorithms are posed as black boxes and can change dynamically over time. Furthermore, sanitizing all the possible principal and proxy variables that could lead to biased decisions can be a formidable task. It would require a fast and efficient rational counterfactual system that intervenes the variables iteratively to find out the antecedent for the given consequent.

Although the proposed framework is most comprehensive in that it can identify a wide spectrum of antecedent variables and values that match a rational counterfactual for a given factual, as with any framework, it has several limitations. The framework, first of all, does not provide any elaborated or implemented rational counterfactual system but we simply assumed one exists. The framework might not be useful in case where algorithms are disclosed. In such a situation, though, it could be used to complement auditing the disclosed algorithms. In addition, the framework is unlikely to find right antecedent variables and values of a rational counterfactual if its base factual statement is constructed from an algorithmic decision that contains an unintentional bias or inadvertent discrimination. The same seems to hold true for an opaque bias or discrimination that is difficult or impossible to detect a priori. Finally, implementing algorithm audits using the framework or any other existing audit study designs can encounter legal resistance (Farivar, 2016).

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016, May 23). *Machine Bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Angwin, A. (2016, August 1). *Make Algorithms Accountable*. The New York Times. https://www.nytimes.com/2016/08/01/opinion/make-algorithms-accountable.html

Asplund, J., Eslami, M., Sundaram, H., Sandvig, C., and Karahalios, K. 2020. "Auditing Race and Gender Discrimination in Online Housing Markets," In *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1), pp. 24-35. https://www.aaai.org/ojs/index.php/ICWSM/article/view/7276

Bandy, J. and Diakopoulos, N. 2020. "Auditing News Curation Systems: A Case Study Examining Algorithmic and Editorial Logic in Apple News," In *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1), pp. 36-47. Retrieved from https://www.aaai.org/ojs/index.php/ICWSM/article/view/7277

Baert, S. 2014. "Wage Subsidies and Hiring Chances for the Disabled: Some Causal Evidence," *The European Journal of Health Economics* (17:1), pp. 71-86.

Baert, S., de Visschere, S., Schoors, K., Vandenberge, D., and Omey, E. 2016. "First Depressed, then Discriminated Against?," *Social Science & Medicine* (170). pp. 247-54.

Bannister, F and Connolly, R. 2011. "The trouble with transparency: A critical review of openness in e-government." *Policy & Internet* (3:1). p. article 8.

Barocas, S., and Selbst, A. D. 2016. "Big data's disparate impact,"*California Law Review* (104), pp. 671-732.

Becker, G. S. (2010). *The Economics of Discrimination*. University of Chicago Press.

Bendick, M., Jackson, C. W, and and Romero, J. H. 1997. "Employment Discrimination Against Older Workers: An Experimental Study of Hiring Practices," *Journal of Aging & Social Policy* (8:4), pp. 25-46.

Bertrand, M., and Mullainathan, S. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *The American Economic Review* (94:4), pp. 991-1013.

Bygren, M., Erlandsson, A., and Gähler, M. 2017. "Do Employers Prefer Fathers? Evidence from a Field Experiment Testing the Gender by Parenthood Interaction Effect on Callbacks to Job Applications," *European Sociological Review* (33:3), pp. 337-48.

Calders, T., and Verwer, S. 2010. "Three naive bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery* (21:2), pp. 277–292.

Cantone, J. A. 2020. "Counterfactual Thinking, Causation, and Covariation in Mock Juror Assessments of Negligence: Twenty-Five Years Later," *Psychological Reports* (123:2), pp. 371-394.

Crawford, K. 2013. "The Hidden Biases in Big Data," *Harvard Business Review* (available at https://hbr.org/2013/04/the-hidden-biases-in-big-data).

Custers, B., Calders, T., Schermer, B., and Zarsky, T. (2012). *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*. Springer Science & Business Media.

Daniel, W. W. (1968). *Racial Discrimination in England*. Baltimore, MD: Penguin Books.
DARPA (July, 2017). *Explainable Artificial Intelligence (XAI)*. DARPA. https://www.darpa.mil/program/explainable-artificial-intelligence

de Bruijn, H., Warnier, M. and Janssen, M. 2022. "The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making." *Government Information Quarterly* (39:2), pp. 1-8.

Deming, D. J., Yuchtman, N., Abulafi, A., Goldin, C., and Katz, L. F. 2016. "The Value of Postsecondary Credentials in the Labor Market: An Experimental Study." *The American Economic Review* (106:3), pp. 778-806.

Diakopoulos, N. 2016. "Accountability in algorithmic decision making," *Communications of the ACM* (59:2), pp. 56-62. DOI: 10.1145/2844110.

Dodge, Y. (2003). *The Oxford Dictionary of Statistical Terms*. Oxford University Press.

Elga, A. 2000. "Statistical Mechanics and the Asymmetry of Counterfactual Dependence," *Philosophy of Science* (68:3), Supplement, pp. 313–24.

Evans, D. N. 2016. "The Effect of Criminal Convictions on Real Estate Agent Decisions in New York City," *Journal of Crime and Justice* (39:3), pp. 363-379.

Farivar, C. 2016. *To study possibly racist algorithms, professors have to sue the US*. Ars Technica. https://arstechnica.com/tech-policy/2016/06/do-housing-jobs-sites-have-racist-algorithms-academics-sue-to-find-out/

Figinski, T. F. 2017. "The Effect of Potential Activations on the Employment of Military Reservists: Evidence from a Field Experiment," *ILR Review* (70:4), pp. 1037-56.

Gaddis, S. M. 2018. "An Introduction to Audit Studies in the Social Sciences," in *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, edited by S. M. Gaddis, 3-44. Springer. doi:10.1007/978-3-319-71153-9

Garfinkel, S., Matthews, J., Shapiro, S. S., and Smith, J. M. 2017. "Toward algorithmic transparency and accountability," *Communications of the ACM* 60(9), pp. 5.

Gell-Redman, M., Visalvanich, N., Crabtree, C., and Fariss, C. J. 2017. "It's All About Race: How State Legislators Respond to Immigrant Constituents." Available at SSRN: https://ssrn.com/abstract=2999173

Goodman, B. W. 2016. "A Step Towards Accountable Algorithms?: Algorithmic Discrimination and the European Union General Data Protection," In *Proceedings of the 29th Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain.

Guszcza, J., Rahwan, I., Bible, W., Cebrian, M., and Katyal, V. 2018. "Why We Need to Audit Algorithms," *Harvard Business Review* (available at https://hbr.org/2018/11/why-we-need-to-audit-algorithms).

Hall, N. (2004). "Two Concepts of Causation", in Collins, J., Hall, N., and Paul, L. A. (eds.), pp. 225–76, *Causation and Counterfactuals*, The MIT Press. DOI: https://doi.org/10.7551/mitpress/1752.001.0001

Hao, K. and Stray, J. 2019. "Can you make AI fairerthan a judge? Play our courtroom algorithm game," *MIT Technology Review* (available at https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/)

Hempel, C. G., and Oppenheim, P. 1948. "Studies in the logic of explanation," *Philosophy of Science* (15:2), pp. 135–175. DOI:10.1086/286983

Hitchcock, C. 2001. "The Intransitivity of Causation Revealed in Equations and Graphs," *Journal of Philosophy* (98:6), pp. 273–99.

Hitchcock, C. 2007. "Prevention, Preemption, and the Principle of Sufficient Reason," *Philosophical Review* (116:4), pp. 495–532.

Kalhan, A. 2013. "Immigration policing and federalism through the lens of technology, surveillance, and privacy," *Ohio State Law Journal* (74:6), pp. 1105-1166.

Levinson, R. M. 1975. "Sex Discrimination and Employment Practices: An Experiment with Unconventional Job Inquiries," *Social Problems* (22:4), pp. 533-43.

Lewis, D. 1973. "Causation," *The Journal of Philosophy* (70:17), Seventieth Annual Meeting of the American Philosophical Association Eastern Division, pp. 556-567.

Lewis, D. 1986. *Philosophical Papers: Volume II*. Oxford: Oxford University Press.

Li, W., Hardesty, D. M., and Craig, A. W. 2018. "The impact of dynamic bundling on price fairness perceptions," *Journal of Retailing and Consumer Services* (40), pp. 201-212.

Marwala, T. (2014). *Artificial Intelligence Techniques for Rational Decision Making*. Springer, Heidelberg.

Marwala, T., and Hurwitz, E. (2017). *Artificial Intelligence and Economic Theory: Skynet in the Market*. DOI: 10.1007/978-3-319-66104-9_12

McCombs, M. 2005. "A Look at Agenda-setting: past, present and future," *Journalism Studies* 6(4), pp. 543–557.

Mellor, D. H. 1995. *The Facts of Causation*, London: Routledge.

Mellor, D. H. 2004. "For Facts as Causes and Effects", in Collins, J., Hall, N., and Paul, L. A. (eds.), pp. 309–24, Causation and Counterfactuals, The MIT Press. DOI: https://doi.org/10.7551/mitpress/1752.001.0001

Menzies, P., and Beebee, H. (2019). Counterfactual Theories of Causation. In *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/causation-counterfactual/

Miller, C. C. (2015, August 10). *Algorithms and Bias: Q. and A. With Cynthia Dwork*. The New York Times. https://www.nytimes.com/2015/08/11/upshot/algorithms-and-bias-q-and-a-with-cynthia-dwork.html

Millican, P. (2007). *David Hume: An Enquiry Concerning Human Understanding*, Oxford University Press: Oxford.

Mishel, E. 2016. "Discrimination Against Queer Women in the U.S. Workforce: A Resume Audit Study," *Socius: Sociological Research for a Dynamic World*. DOI: 10.1177/2378023115621316

Paul, L. A., and Hall, N. (2013). *Causation: A User's Guide*, Oxford: Oxford University Press.

Pearl, J. (2000). *Causality*, Cambridge: Cambridge University Press.

Raghavan, M., Barocas, S., Kleinberg, M., and Levy, K. 2020. "Mitigating bias in algorithmic hiring: evaluating claims and practices," In *Proceedings of the 2020 Conference on Fairness, Accountability, and TransparencyJanuary*, pp. 469–481. DOI: 10.1145/3351095.3372828

Rivera, L. A. and Tilcsik, A. 2016. "Class Advantage, Commitment Penalty: The Gendered Effect of Social Class Signals in an Elite Labor Market," *American Sociological Review* (81:6), pp. 1097-1131.

Ruffle, B. J. and Shtudiner, Z.. 2015. "Are Good-Looking People More Employable?," *Management Science* (61:8), pp. 1760-1776.

Salvatier J., Dafoe A., Zhang B., and Evans O. 2018. "When Will AI Exceed Human Performance? Evidence from AI Experts". arXiv:1705.08807v3 [cs.AI].

Sandvig, C., Hamilton, K., Karahalios, K., and Langbort, C. 2014. "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms," presented to *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*, a preconference at the 64th Annual Meeting of the International Communication Association. May 22, 2014; Seattle, WA, USA.

Shoemaker, P. J., and Vos, T. P. (2009). *Gatekeeping Theory*. New York: Routledge.

Simon, H. A. (1957). *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*, second edition, New York: Macmillan.

Smith, A. 2018. *Public Attitudes Toward Computer Algorithms*. Pew Research Center. https://www.pewresearch.org/internet/2018/11/16/public-attitudes-toward-computer-algorithms/

Snow, J. (April 9, 2022). *Cities Take the Lead in Setting Rules Around How AI Is Used*. Wall Street Journal. https://www.wsj.com/articles/cities-take-lead-setting-rules-around-how-ai-is-used-11649448031

Starr, W. (2019). Counterfactuals. In *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/counterfactuals/

Vijayakumar, S. 2017. "Algorithmic Decision-Making," *Harvard Political Review* (44:2), pp. 9-11.

Wallace, M., Wright, B. R. E, and Hyde, A. 2014. "Religious Affiliation and Hiring Discrimination in the American South: A Field Experiment," *Social Currents* (1:2), pp. 189-207.

Wang R., Harper F. M., and Zhu H. 2020. "Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences," In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, April 2020 pp. 1-14. DOI: 10.1145/3313831.3376813

Woodward, J. 1996. "Explanation, Invariance, and Intervention," *Proceedings of the 1996 Biennial Meetings of the Philosophy of Science Association*. Part II: Symposia Papers, pp. S26-S41.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*, Oxford: Oxford University Press.

Woodward, J. and Hitchcock, C. 2003. "Explanatory generalizations. Part I: a counterfactual account," *Noûs* (37:1), pp. 1–24.