EXPLAINABLE AI
as a tool for risk managers

December 2021

Professional Risk Managers' International Association

PRMIA INSTITUTE
Professional Risk Managers' International Association

Author

Hersh Shefrin, Mario L. Belotti Professor of Finance at Santa Clara University

Hersh is one of the pioneers in the behavioral approach to economics and finance. Together with Nobel laureate Richard Thaler, he developed an economic theory of self-control. The January 2001 issue of CFO magazine lists him among the academic stars of finance. A 2003 article in the American Economic Review listed him as one of the top fifteen economic theorists to have influenced empirical work.

He is the author of seven books and numerous articles about how behavioral ideas impact real world decisions. In 2009, his book Beyond Greed and Fear was recognized by J.P. Morgan Chase as one of the top ten books published since 2000. He received his B.Sc. (Hons.) degree from the University of Manitoba, an M.Math degree from the University of Waterloo, and his Ph.D. from the London School of Economics. He also holds an honorary doctorate from the University of Oulu, Finland.

He is frequently interviewed by the press and his work was profiled by BBC-TV in February 2014. He writes a monthly blog post for Forbes, and has intermittently written for The Wall Street Journal, The Huffington Post, and VOX. His Twitter handle is @HershShefrin.*

Editor

David M. Rowe, Founder and president of David M. Rowe Risk Advisory

David M. Rowe is founder and president of David M. Rowe Risk Advisory, a risk management consulting firm. Dr. Rowe has spent over 40 years in the risk management technology, banking and economic forecasting industries. He authored the monthly Risk Analysis column in Risk magazine from 1999 to 2015. He also is the author of the recently published book An Insider's Guide to Risk Management – Relearning the Lessons of the Global Financial Crisis.

* I am grateful to Sanjiv Das, Krishna Gade, Chris Russell, and David Rowe for their valuable comments on this paper. I am also grateful to the participants at the Herbert Simon Society International Workshop "The Digital World, Cognition and Behavior" which took place in Turin, Italy in November 2019, where I presented an early version of this work with the title "Explainable AI and Bounded Rationality."

About PRMIA Institute

PRMIA Institute – Industry Nexus for Risk Standards

The PRMIA Institute serves the global risk management community with leading thought, peer-vetted research, and stewardship of the risk management profession.

We help your risk organization navigate complexity and aim to:

- Advance thought leadership in risk management and develop standards of practice for risk management.
- Promote understanding of the field, both from a practice and policy perspective.
- Work with private and public entities on applied risk management related research.

Our work is centered on applied research.

We partner with individual organizations as well as multilateral funding initiatives.

- The PRMIA Institute is a registered party for the EU's Horizon 2020 funding initiative.
- We have experience in design and management of work packages of larger funding initiatives.
- Our PRMIA expert network and the PRMIA Institute senior advisors are industry leaders with vast experience.

All this allows the Institute to be a research partner with the knowledge and network to bring success to your analysis.

Table of contents

Executive summary

As context, a 2020 survey conducted by the World Economic Forum found that over 85% of financial services firms have implemented some form of artificial intelligence (AI).[1] Forecasts over the current decade suggest that the global business value of AI in banking might reach $300B.[2] Plausibly, the use of AI-based applications will be a primary driver of growth in the banking sector, especially in areas such as credit lending, fraud and threat detection, anti-money laundering, and compliance.

Improvements to machine learning (ML) technology are taking place rapidly, and while finding widespread application, are "black box" in design, which obscures how inputs are converted into outputs. The black box design can be especially problematic when algorithms generated by machine learning produce outcomes that appear to be unfair in the sense of exhibiting bias.

ML, a special case of artificial intelligence, is changing the nature of risk management, especially in financial services. Traditional risk management has relied on an approach known as Model Risk Management (MRM).[3] However, MRM was designed for models which predate and are less complex than the models used in ML.

Because of the rapid growth in ML techniques, some risk management groups are restructuring themselves to deal with a different set of challenges than they faced in the past. In particular, some MRM teams now include a member with a background in data science, called an AI validator.[4] At the moment, this approach will work in some instances, mostly within large firms, but is not realistic for all firms. Indeed, many firms will choose to rely on consultants and external validators.

Biased outcomes from ML-generated algorithms with black box features create risk for the users of these algorithms, especially in situations where disadvantaged parties seek legal redress.[5] Data scientists have been responding to issues of algorithmic bias in several ways, including developing tools for explainable AI, which seek to shed light on the inner workings of ML-generated algorithms. Explainable AI is a relatively new approach which analyzes ML-generated decision models and produces output which can serve as documentation for various constituencies, including regulatory bodies. Explainable AI is an important tool for AI validators and is being developed by large firms such as Amazon, IBM, Google and Microsoft, as well as small firms such as Fiddler AI, Parity, Weights and Biases, and ZestFinance.

At present, the field of explainable AI is in flux and has not yet produced a definitive approach for analyzing the nature of ML-generated algorithms. Indeed, there is a realization in the literature on explainable AI about a lack of "robustness" (or "stability").[6] At the heart of the robustness issue is the extent to which explainable AI models actually shed light on causal mechanisms in the real world, or instead are simply effective at predicting the input-output relationship in ML-generated algorithms.

In particular, risk managers need to be clear about the questions to which they seek answers from explainable AI. Most ML-generated algorithms rely on large amounts of data, and it can be easy to lose sight of the forest for the trees in these settings. Risk managers who use or are planning to use explainable AI can benefit from examining small, pertinent examples that allow them to follow exactly how specific explainable AI techniques work, to help themselves develop an intuitive understanding of explainable processes.

In 2017, two academics at the University of Washington, Scott Lundberg and Su-In Lee, applied a concept from game theory known as the Shapley value, to develop an explainable AI technique they called SHAP. For the moment, think of SHAP as an approach for analyzing the relative impacts of different features in an ML-generated algorithm.[7]

Notably, the history of SHAP illustrates some of the main messages for risk managers about ML and explainable AI, especially in respect to subtleties, counterintuitive elements, the importance of being clear about the questions risk managers are seeking, and the importance of being aware of what kinds of answers explainable AI techniques can provide. In addition, explainable AI can be opaque and exhibit bias, so users of explainable AI need to exercise caution.

SHAP has come to be used by practitioners in financial services, especially to improve processes related to lending and credit management. Some SHAP users discovered that the technique has flaws, and so developed alternative approaches. In 2019, during Congressional testimony about the growing role of AI in finance, mention was made of SHAP, its associated flaws, and efforts to address them. Through 2019 and 2020, data scientists worked on these issues, some by making relatively modest modifications to the original SHAP approach. These efforts are still underway, and no definitive approach has yet to emerge.

The efforts just mentioned have cast light on some counterintuitive behavioral issues. One important issue involves the concept of “less is more,” which in this case pertains to the discarding of distributional information. A related issue is the tendency to conflate the character of algorithmic bias with the extent of its presence in a population. These are separate concepts, and care needs to be taken not to confuse their effects. Indeed, conflating the two is a strong contributor to the reason why “less is more” has made sense in developing explainable AI techniques.

Bias and lack of fairness are central issues in ML research. In 2021, data scientists began to focus on how the types of fairness issues they have been studying compare to the treatments of fairness and discrimination in legal and regulatory settings. Their efforts have focused on statistical techniques to measure what has come to be called “demographic disparity” for detecting biases: these techniques provide perspectives which are consistent with the manner in which the law approaches issues of bias. It will be important for risk managers to understand the complementary relationship between the demographic disparity approach and the approach to explainable AI embodied in techniques like SHAP.

From both business and legal perspectives, one of the most important challenges risk managers face as a result of using black box technologies involves communicating with other parties about both risks and rationales for why particular past decisions were made. In this regard, explainable AI will be an important tool for risk managers, and being able to communicate effectively requires having as good an understanding as possible about how the black boxes are "thinking." To this end, risk managers need to have a clear sense of the questions to which they seek answers, so that they can select explainable AI techniques which are most likely to produce sensible answers to these questions.

In summary, here are the main messages in the paper.

- The field of explainable AI is in flux and has not yet produced a definitive approach for analyzing the nature of ML-generated algorithms.

- Explainable AI, while using surrogate models to mimic the input-output relationship in ML-based algorithms, might not capture the underlying real world process, and therefore lack robustness in respect to the degree of explainability.

- Both ML-generated algorithms and explainable AI involve important subtleties and counterintuitive elements which are important for users to understand. One of the most counterintuitive elements is that ignoring valid distributional information can be valuable for explainability.

- From both business and legal perspectives, one of the most important challenges risk managers face as a result of using black box technologies involves communicating with other parties about risks and rationales for why particular past decisions were made. To address these challenges, risk managers will need to use explainable AI technology.

- The shift to ML-generated algorithms generates risk management challenges, in respect to the management of data lifecycle issues. ML-based model lifecycle management will be necessary to address dynamic issues, such as the value of specific models declining with the arrival of new data and the change in task environments. The management of lifecycle risk will involve choices about associated software products,[8] differential skills, associated division of tasks, the flow of work, and traditional silo issues related to information sharing.

- Risk management is a quantitatively oriented discipline. Risk managers will face important challenges as they deal with lifecycle issues associated with ML-based modeling, in the course of engaging in monitoring, applying explainable AI techniques, and addressing biases which these techniques uncover. Risk managers can better understand and communicate risks associated with ML-generated algorithms and explainable AI by investing time and effort to work through small numerical examples.[9]

1. Introduction

Advances in data science and computing hardware have led to innovations in artificial intelligence (AI) which are revolutionizing decision making across the globe. This revolution has brought both new opportunities and new risks. An important source of risk stems from the character of machine learning (ML), which by its nature involves a set of opaque processes, aptly described as black boxes.

Black boxes, when used in high stake situations, are vulnerable to high risks because when things go wrong, the source of the problems they generate can be difficult to diagnose and therefore to correct easily. In this paper, I describe the general nature of the challenges posed by reliance on AI systems which are opaque, and correspondingly the character of approaches for dealing with the associated opacity.

Consider an example involving two related but distinct issues, the first being ML-based classification technology and the second being potential biased classifications.[10] Apple is one of the world's most valuable companies. One of its business lines involves consumer finance, the processing of consumers' financial transactions and associated credit. Apple uses ML classification technology to offer its services to prospective customers. Stephen Wozniak, together with the late Steve Jobs, co-founded Apple. In 2019, Wozniak reported that Apple Card separately offered him and his wife digital credit lines. Notably, Wozniak and his wife hold all of their assets jointly. Nevertheless, Apple Card offered Wozniak a credit line with a limit that was ten times greater than the corresponding limit offered to his wife. The Wozniaks were baffled by the asymmetric treatment, which some might interpret as exhibiting bias in the sense of being discriminatory and unfair.

The Wozniaks were not alone in receiving asymmetric treatment. They were not alone in being baffled by the asymmetry, and at the time Apple Card did not offer an explanation that suggested they understood why their algorithms generated such disparate offers. Certainly, the recipients of Apple Card's offer who noted the asymmetry did not understand.

The Apple Card example involves reliance on an ML-based algorithm for classifying customers, which produces outputs featuring apparent gender bias. To gain some perspective on the general issues under discussion, consider that the human brain is also something of a black box which engages in classification. Facial recognition is a classification task. People routinely process visual inputs as they classify people's faces into categories defined, for example, by gender, age, ethnicity, and family membership.

Although humans are generally quite good at facial recognition, very little is known about how human brains are able to process the inputs as successfully as they do. Moreover, the human brain is not infallible. Some people have difficulty with facial recognition due to a condition known as "visual agnosia." This condition was highlighted in a book authored by the late neurologist Oliver Sacks, which he titled "The Man Who Mistook his Wife for a Hat." As a result of suffering from visual agnosia, the man in question, I will call him Mr. WH," did indeed mistakenly classify his wife's head as his hat, and at the conclusion of a medical appointment with Dr. Sacks, Mr. WH reached to put "it" on.

The man who mistook his wife for a hat
by Oliver Sacks

The “wife-hat case” provides important insights for bias in AI classification systems. For Mr. WH, visual agnosia came on slowly, and as the disease progressed, the man learned to adapt. Notably, Mr. WH recognized some aspects of the data he was observing. For instance, he could recognize visual inputs such as shapes, color, and textures. He could also associate other sensory data, such as voices and odors, to objects he was seeing. However, what Mr. WH’s brain could not do was to assemble the data to produce a classification judgment such as “this is my hat,” “this is my wife,” “this is my foot,” or “this is my shoe.” His internal classification algorithm was impaired.

Especially relevant to understanding AI bias is what Mr. WH did to circumvent his classification impairment. He relied on patterns based on sensory cues that were from nonvisual sources, such as speaking voices and smells, to provide him with a clue about the identity of the person with whom he was interacting.

This pattern recognition technique worked well for him in familiar settings, and apart from his wife, most people who knew him were unaware of his impairment. However, non-familiar settings presented a problem, especially if his cues were obfuscated. For example, in a room with many people speaking simultaneously, Mr. WH might be unable to pick out the speaking voice of a specific person. Or Mr. WH might be familiar with the person speaking but not singing and be in a situation where that person is singing. Or the person might speak English with a foreign accent which sounds to Mr. WH very similar to the voices of many other people who speak with a similar accent.

Machine learning performs classification tasks in ways that are similar to the way that Mr. WH deals with visual agnosia. Machines arrive at classifying algorithms by identifying patterns and picking up on cues. In situations with which there is ample input data (familiarity), and the quality of the input data is high, machine learning algorithms often work well. However, in other types of situations, machine algorithms might make classification errors such as mistaking Mr. WH’s wife for a hat.[11] Intrinsically, the machine does not understand the essence of Mr. WH’s wife in the way as a healthy person understands the essence of Mr. WH’s wife. Instead, the algorithm focuses on cues relating to Mr. WH’s wife in the same way as does the impaired Mr. WH.

Mr. WH's disability prevented him from processing facial information in the same way as normal people. Consider an example involving potential information processing bias, based on a real-world situation, in which students are applying for admission to a top university. Suppose half the applicants are male and half are female, but the university's admission algorithm rejects more females than it does males. Is this discrimination? In the real-world case, which involved the University of California, Berkeley, the asymmetric acceptance rates stemmed from the fact that females tended to apply to programs that had especially high admission standards. Imagine that these programs are so selective that most applicants are denied admission.[12] If mostly females apply to these programs, mostly female applicants will be rejected, even if every university program features an unbiased admissions policy. In fact, if an ML algorithm is trained on these data, but without data on programs being applied for, it might well produce a biased admissions algorithm. Just as Mr. WH could not process facial information, the algorithm might not be able to process information about the type of program involved in the application, perhaps because these data are not available.

Regarding facial recognition and algorithms, studies show that facial recognition services can be biased against women and people of color when they are trained on photo collections dominated by white men.[13] As reported in the New York Times, there are three types of biases associated with machine learning: biases in the data, biases in the algorithms, and biases in the humans that use them.

Machine learning biases occur in many domains and can be quite pronounced. For this reason, their use has been attracting the attention of legislators and regulators. This is certainly the case with FinTech. In May 2019, the US House Committee on Financial Services announced the creation of two task forces, one on FinTech and one on AI.[14] While the Committee structured the task force on FinTech to have a near-term focus on applications such as quickly instituting new regulations for activities such as underwriting and payments, it structured the task force on AI to focus on long-term risk issues such as those involving fraud and digital automation.

In April 2020, the Federal Trade Commission issued a warning about using AI systems that might lead to decisions featuring racial bias and result in some individuals being discriminated against in respect to employment, housing, and insurance. During the same month, the European Union released draft regulations that described sanctions for companies offering such technology.[15]

As for Apple Card, in 2020, New York State regulators launched an investigation into the issue out of a concern that it discriminated against women. Goldman Sachs operates Apple Card. As it happens, in the end, these regulators ruled that in this particular case, Goldman Sachs did not engage in discriminatory behavior.[16] The New York State Department of Financial Services stated that although some Apple Card customers might have been authorized users on their spouses' accounts, this was different from being the primary account holder. More importantly, the Department concluded that this distinction was germane to their finding that Goldman Sachs' operation of Apple Card was nondiscriminatory. The Department added that Apple Card could have been more transparent with its customers about the criteria it used when making offers. Of course, given its black box ML approach, doing so would require explainable AI technology.

The regulatory issues are broad. Many of the situations which feature algorithmic bias involve classification tasks. For example, algorithms are often used to classify loan applicants by creditworthiness and job applicants by ability.

In respect to the first issue, loan applications, consider the following: When lenders use AI to classify loan applicants into two subgroups, accept and reject, they face a legal issue. This is because when a consumer is denied credit, the Fair Credit Reporting Act of 1970 requires accurate and actionable reasons for the decision so that consumers can repair their credit and re-apply successfully.

In respect to practices including hiring, consider the following: In April 2020, the Federal Trade Commission issued a warning about using AI systems that might lead to decisions featuring racial bias and result in some individuals being discriminated against in respect to employment, housing, and insurance. During the same month, the European Union released draft regulations that described sanctions for companies that offered such technology.[17]

Algorithmic classification typically entails the analysis of a myriad of characteristics to search for patterns in the data that serve to assign objects to different classes such as financially sound or unsound, high quality or low quality, and high risk or low risk. Bias stems from classification rules which on average produce erroneous assignments.

"Explainable AI" is a term used to describe techniques for generating surrogate models to explain the manner in which ML black box algorithms transform inputs into outputs. One can think of explainable AI as a class of risk management techniques, designed to help users of these algorithms assess their behavior before they apply them in practice, as well as during the time they are applied in practice.[18]

The June 2019 hearing of the US House Committee on Financial Services featured a discussion about explainable AI. In this regard, the CEO of the firm ZestFinance stated that lenders use his firm's software "to make their lending fairer."[19] He went on to say the following: "There are purported to be a variety of methods for understanding how ML models make decisions. Most don't actually work. As explained in our white paper and recent essay on a technique called SHAP, both of which I've submitted for the record, many explainability techniques are inconsistent, inaccurate, computationally expensive, or fail to spot discriminatory outcomes."[20]

The technique called SHAP, mentioned in the previous paragraph, is important. In 2017, two academics at the University of Washington, Scott Lundberg and Su-In Lee, applied a concept from game theory known as the Shapley value, to develop an explainable AI technique they called SHAP, an acronym for Shapley additive explanations.[21] For the moment, think of SHAP as a surrogate function that mimics the black box ML function by assigning local relative attribution weights to different variables in the black box multivariate function. In the last sentence, the term "local" means that the attribution weights can vary at different points in the function domain. Notably, the presence of local variation, or unstable attribution assignment, can be construed as a lack of robustness in the provision of explanations.[22]

In this paper, I employ SHAP as a vehicle to illustrate how explainable AI works.[23] For this purpose, I use simple examples, analyzed in the explainable AI literature, in order to make the exposition as transparent as possible.[24] Although the discussion in the body of the paper will be narrowly focused for the purpose of exposition, the issues raised are quite general.

Part of the discussion about SHAP will focus on its flaws. This is intentional because understanding the nature of these flaws provides insight into how using both ML-generated algorithms and explainable AI technology, while valuable, expose users and other parties to risks.

The examples, while simple, are detailed. This is important because the insights about when explainable AI techniques do work and do not work lie within the details. This is why detailed examples lie at the heart of the paper. It is only by working through simple examples that risk managers can understand how explainable AI works, why it can produce flawed results, and how to remedy these flaws.

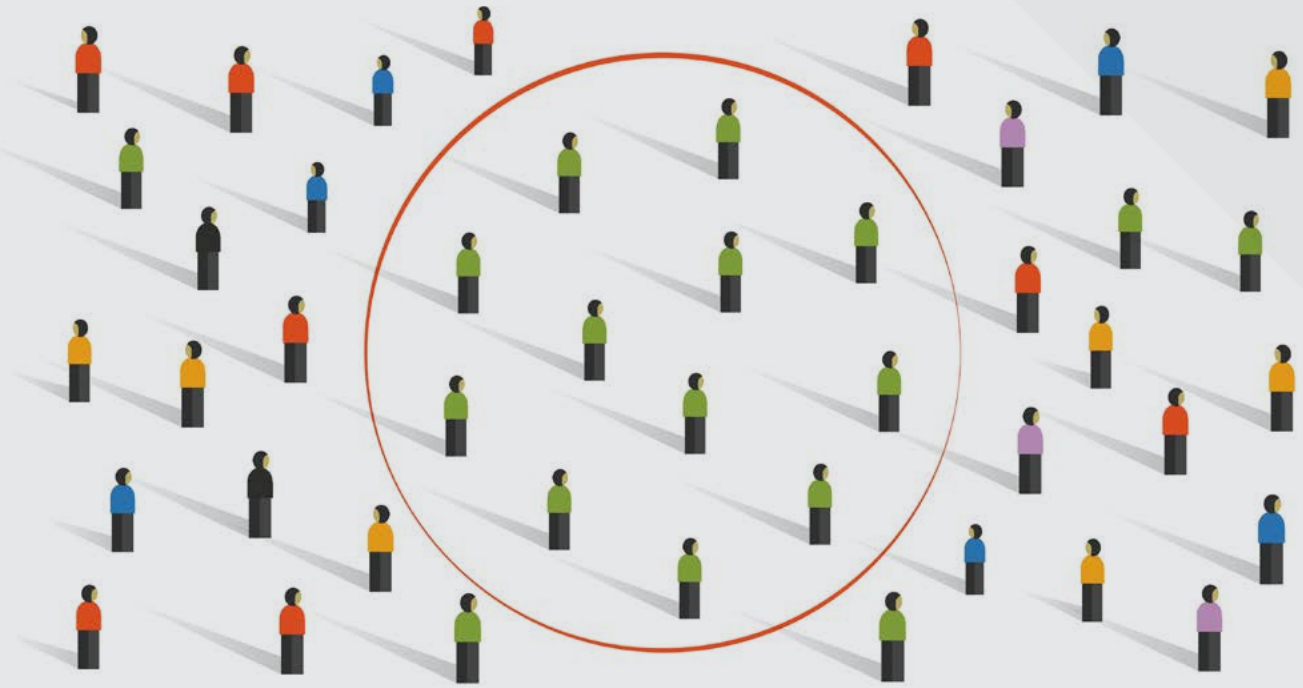
In summary, the main points of the paper are as follows: The field of explainable AI is in flux and has not yet produced a definitive approach for analyzing the nature of ML-generated algorithms. Not all explainable AI are robust. Both ML-generated algorithms and explainable AI involve important subtleties and counterintuitive elements which are important for users to understand. From both business and legal perspectives, one of the most important challenges risk managers face as a result of using black box technologies involves communicating with other parties about risks and rationales for why particular past decisions were made. To address these challenges, risk managers will need to use explainable AI. Risk management is a quantitatively oriented discipline, and risk managers can better understand and communicate risks associated with ML-generated algorithms and explainable AI by making the investment of time and effort to work through small numerical examples.

2. A stylized example of analyzing fairness in the performance of a classification task

In this section I introduce two simple stylized examples for discussing the two key distinct themes of this paper, explainable AI and fairness risk. The examples, while differing in institutional context and distributional features, have the same underlying formal structure.

The first example relates to financial services, such as classifying loan applicants by acceptance or rejection. In the example, loan applicants are characterized by two features, gender (male, female) and net worth (low, high). Imagine that there are 100 applicants of which 70 are male with high net worth, 10 are male with low net worth, 10 are female with high net worth, and 10 are female with low net worth. The loan application policy recommends that loans only be granted to male applicants with high net worth.[25] Think of the loan application policy as being ML-generated.[26]

The second example relates to recruitment. In the example, a hiring manager who is male interviews applicants for jobs at a furniture moving company. Some jobs involve physically lifting furniture, and some jobs provide logistical support. I will present several versions of the example. Each version involves a hiring manager who uses a policy to classify applicants as either hired or rejected, based on input about two features: gender (male or female) and whether or not the applicant has lifting ability. For now, imagine that in this example 100 people have applied, with 50 of the applicants being male who have lifting ability, 40 being males with no lifting ability, and 10 being female with no lifting ability. Think of the hiring policy as being ML generated.



Intuitively, think about the kinds of policies that would be considered fair to applicants. In the loan application example, a fair acceptance policy might be one that treats gender as irrelevant. This would be in line with the Apple Card situation discussed above. Accepting all loan applicants would qualify as fair, or accepting any and all high net worth applicants would qualify as fair.

The hiring example is a bit more complicated because the moving company hires people to do both physical tasks and logistical tasks.[27] If the hiring manager hired every applicant, that would seem to be fair, as it discriminates against no one. However, what if the hiring manager were to classify as hired all 50 applicants with lifting ability, 20 male applicants without lifting ability, and 5 of the 10 female applicants without lifting ability? Does this policy appear to discriminate against females? The hiring manager might argue not, pointing out that in selecting the 75 applicants classified as hired, he/she ranked candidates first on lifting ability, and for those without lifting ability, hired half the males in the applicant pool and half the females in the applicant pool.

There is a criterion known as "conditional demographic disparity" (CDD) which can formally test for the presence of gender bias.[28] In the moving company example, the condition involves comparing the percentage of all those hired who are female to the percentage of all those rejected who are female, first among those without lifting ability and then among those with lifting ability. CDD asks for each subgroup separately, whether among those rejected, females are disproportionately represented than they are among those who are hired? If the answer turns out to be yes for both subgroups, then CDD signals that female applicants are victims of gender bias.[29]

For applicants without lifting ability, females comprise 20% (= 5/25) of those who are hired, and 20% of those rejected. The CDD test signals discriminatory practices when the probability associated with rejection exceeds the probability associated with acceptance. In this example, the two probabilities are equal, and therefore the CDD criterion does not signal gender bias. As no female applicants have lifting ability, for applicants with lifting ability, females hired and rejected make up 0% of all of those with lifting ability who are hired, but the corresponding ratio for those rejected is undefined (= 0/0).

Next, turn to a version of the example which might be construed as unfair: What if the hiring manager classified as hired the same 50 applicants with lifting ability, but now all 40 male applicants without lifting ability, and the same 5 female applicants without lifting ability? CDD would raise a red flag in this case: rejected female applicants without lifting ability now comprise 100% of the rejected applicants with no lifting ability, but female applicants without lifting ability only comprise 5.5% (= 5/45) of hires from this subgroup. The CDD red flag stems from the inequality 100% > 5.5%. (CDD also requires a similar test for applicants with lifting ability. However, with there being no female applicants with lifting power, the corresponding ratio comparison has little meaning.)

To round out the discussion, consider the loan application example in respect to CDD. For high net worth applicants, females comprise 100% of those rejected and 0% of those accepted. CDD would raise a red flag in this case, based on the inequality 100% > 0%. For low net worth applicants, females comprise two thirds of those rejected. None are accepted, and so the corresponding ratio pertaining to female applicants is undefined.

In closing this section, let me mention that CDD provides the key to understanding why situations such as that described in the introductory section involving college admission policies at UC Berkeley were not biases, despite initial appearances to the contrary. Notably, UC Berkeley situation is not unique, but is representative of a class of situations associated with a phenomenon in statistics known as "Simpson's Paradox."[30]

3. A simple example to explain how SHAP works

In this section, I use the stylized moving company example to show first how SHAP works, and second how SHAP can be used to assess the degree of bias in respect to fairness.

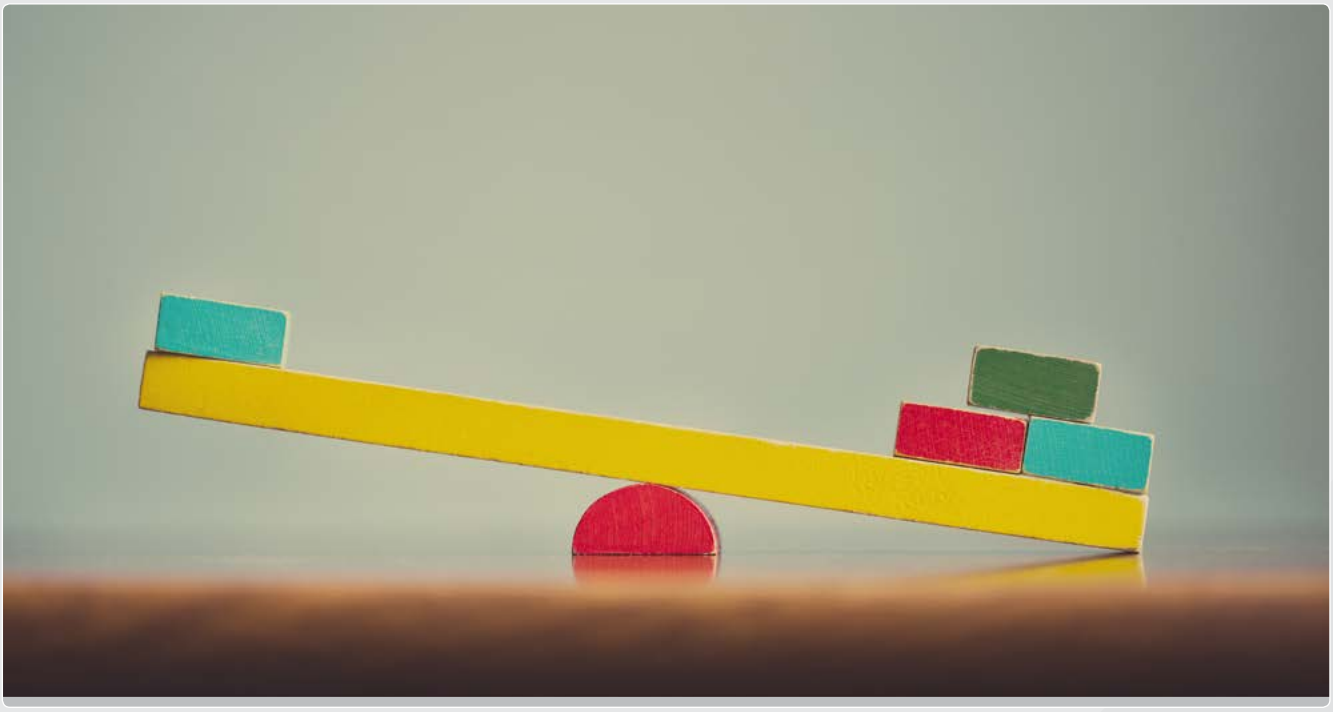
AI validators use explainable AI techniques such as SHAP to provide information about the relative impact of different features in determining ML-generated outcomes. This section focuses on the following two questions about SHAP. What information does SHAP provide? What does SHAP do to produce the information? To answer these questions, the discussion continues with the example introduced in the previous section in which the manager's policy classifies as hired 50 applicants with lifting ability, 40 male applicants without lifting ability, and 5 female applicants without lifting ability.

Consider the question: What information does SHAP provide? In a sentence, SHAP explains how the degree to which different features explain why the hiring probability for a particular applicant subgroup differs from the population average. In this version of the example, 95% of applicants are classified as hired. Call the 95% hiring rate the "baseline." Notably, this version of the model involves only 50% of female applicants with no lifting ability being classified as hired. SHAP explains the degree to which the change in probability from baseline, -45% (= 0.5 – 0.95), is attributable to gender. Specifically, SHAP will decompose the -45% into two portions, one for gender and the other for lifting ability.

As explained below, SHAP will stipulate that the portion attributable to gender is -42.5%, with the residual -2.5% (= -0.45 – (-0.425)) being attributable to lifting ability. Therefore, for this example, SHAP explains that gender contributes 17 times (= 0.425/0.025) as much as lifting ability to the hiring probability for female applicants with lifting ability being 45% below the 95% population average. This would seem to suggest that female applicants are facing discrimination because of gender.

For male applicants with lifting ability, SHAP will attribute half of the incremental 5% hiring probability to gender and the other half to lifting ability.

For male applicants without lifting ability, SHAP will attribute 7.5% of the incremental 5% hiring probability to gender and -2.5% to lifting ability. The fact that the contribution of gender exceeds the whole of incremental hiring probability can be considered a signal from SHAP that male applicants are receiving favored treatment relative to female applicants.



Consider next the question: What does SHAP do to produce this information? Intuitively, SHAP employs Shapley's insight about how to divide the "value pie" in a cooperative game in which players can form coalitions. Shapley argued that each player's fair share from the game can be obtained by listing all the coalitions an individual player might join, and averaging that player's value added from joining a coalition, across coalitions. SHAP treats features as being analogous to players.

Technically, SHAP follows a mechanical protocol as a series of steps, described below as a series of 11 queries and associated answers. The steps are not difficult, just a bit tedious, akin to completing a tax return. Nonetheless, it is important that risk managers understand the protocol. Here are the 11 steps.

1. What percentage of applicants are classified as hired under the policy? The answer is 95%.

2. What is the probability that a male applicant will be classified as hired, averaging across those with lifting ability and those without lifting ability? The answer is 100%, since the policy prescribes that all males be hired.

3. What is the probability that a female applicant will be classified as hired, averaging across those with lifting ability and those without lifting ability? The answer is 50%, since the algorithm prescribes that 50% of females without lifting ability be hired, and there are no females with lifting ability.

4. What is the probability that an applicant with lifting ability will be classified as hired, averaging across males and females? The answer is 100%, as the policy prescribes hiring all of the 50% of the applicant pool consisting of males with lifting ability; and no females have lifting ability.

5. What is the probability that an applicant without lifting ability will be classified as hired, averaging across males and females? The answer is 90%, as 40 of the 50 applicants without lifting ability are male, and the policy prescribes that these 40 males along with 5 females be hired.

   With these answers in hand, SHAP frames the outcomes as incremental hiring probabilities relative to the general average. For instance, as discussed earlier, for the feature "female" the incremental hiring probability is -45%. SHAP continues its protocol of queries as follows, focusing on female applicants with no lifting ability.

6. With what probability does the hiring policy classify a female applicant as hired, averaging across the features "has lifting ability" and "does not have lifting ability?" The answer is 50%: as mentioned earlier, half of female applicants are hired under the policy.

7. With what probability does the hiring policy classify an applicant without lifting ability as hired, averaging across the features male and female? The answer is 90%, as the policy recommends that all 40 male applicants without lifting ability are hired, along with 5 females without lifting ability.

   SHAP next considers a series of incremental changes to hiring probability, as information about features is presented sequentially. The starting probability is the average, 95%.

8. When information about gender is presented first, the hiring probability for females, 50% from step 6, represents a change of -45% from the 95% starting probability.

9. When information about lifting ability is presented next, the hiring probability, now for female applicants with no lifting ability, remains at 50%, a further change of 0% (from the -45% in step 8).

10. Change the order of presentation from gender first and lifting ability second to lifting ability first and gender second. In this case, the hiring probability for no lifting ability is 90% (from step 7), a change of 5% from the 95% average. The hiring probability for female applicants with no lifting ability is 50%, representing a change of -40% from 90%.

11. In respect to steps 8 through 10, SHAP averages the incremental probabilities associated with the two features across the two orders, as order is arbitrary. For gender, the average (of -45% and -40%) is 42.5%. For lifting ability, the average (of 0% and -5% is -2.5%).

4. Flaws in the SHAP protocol

The CDD framework provides criteria to test for the presence of bias in classification algorithms. Correspondingly, attribution frameworks like SHAP seek to explain how the relative contribution of different features to outcomes varies across the objects being classified.

As was mentioned above, in Congressional testimony, the CEO of ZestFinance stated that "many explainability techniques are inconsistent, inaccurate, computationally expensive, or fail to spot discriminatory outcomes," and explicitly mentioned SHAP.[31] Research by data scientists at the firm Fiddler AI sheds light on the nature of what makes SHAP flawed and offers suggestions about improvements.

The issues are subtle. Readers would be hard pressed to find fault with SHAP's attributions in the example discussed in the previous section. In that example, which CDD flags in connection with gender discrimination, SHAP attributes to gender much of the explanation for why the hiring probability for females is lower than average, and why the hiring probability for male applicants without lifting ability is above average. What might be unclear is what the exact attribution weights mean, for example why gender carries 17 times the weight of lifting ability for female applicants.[32]

Work by data scientists at Fiddler AI has helped to articulate the source of the flaws in SHAP, by focusing on examples of blatant bias. For example, suppose that the hiring policy stipulated that all male applicants be hired and all female applicants be rejected, a situation which CDD would clearly identify as being discriminatory. Intuitively, we would expect SHAP to assign 100% of the attribution weight to gender and 0% to lifting ability. However, not so; according to SHAP, gender receives 94.4% of the weight and lifting ability receives 5.6%. For male applicants without lifting ability, gender receives 150% of the weight and lifting ability receives -50%.[33]

The output from SHAP does not concur with what intuition suggests. Is one correct and the other not? Can both be right? The answer is that the version of SHAP described above is flawed. In their original work, Lundberg and Lee made this point and suggested that users of SHAP assume that all features are independently distributed, and use only information about marginal distributions. Fiddler AI data scientists make this clear by analyzing how SHAP would perform in the hiring example discussed above.

In a modified version of the example involving independently distributed features, 90% of the applicants are male, and 50% of the applicants have lifting ability, as is the case in the original data. Given independence, there are 45 male applicants with lifting ability, 45 male applicants without lifting ability, 5 female applicants with lifting ability, and 5 female applicants without lifting ability.[34]

For the version of the example with independently distributed distributions, and the hiring rule in which all male applicants are classified as hired and all female applicants are classified as rejected, what does SHAP do? For all applicants, SHAP attributes 0% weight to lifting ability and 100% weight to gender, which accords with intuition.[35]

It is very important to understand the reason why SHAP produces a result that accords with intuition in the second example featuring independence, but not in the first. The answer lies in the weeds of the 11-step protocol, and I want to alert readers at this point that the issue is subtle and obscure, but critical.

When we delve into the 11-step protocol, here is what happens. The crucial steps are 1, 7 and 10. For step 1, imposing independence does not alter the general hiring rate for the population as a whole. For step 7, there is a difference, and the difference between the two examples, with and without independence, is that the conditional hiring probability for applicants who have lifting ability is 90% with independence and 80% without independence. This difference is due to the presence of female applicants with lifting ability in the example involving independence. Keep in mind that the hiring probability for the general applicant pool is 90% in both versions of the example. Therefore, when executing step 10 for female applicants without lifting ability, when lifting ability appears first in the order, the incremental hiring probability is 0% with independence and -10% without independence. The crucial issue here is that the feature lifting ability needs to have 0% incremental probability when it appears before gender in the protocol.

The computations can be a bit mind numbing, and it is easy for eyes to glaze over in going through the details. To provide a bit more intuition, consider an informal way to think about what is going on. Metaphorically, in the version of the example without independence, it is as if SHAP mistook Mr. WH's wife for a hat. The mistake occurs when SHAP treats the incremental probability associated with lifting ability as reflecting cause and effect. In the version of the example without independence, when SHAP does attribution analysis for female applicants and focuses first on lack of lifting ability, it recognizes that some applicants with lifting ability do get classified as hired. However, it misses the fact that all of those so classified are male. In terms of wives and hats, it is as if SHAP mistakenly identifies some of the males as females.[36]

Readers might recall that in the first version of the furniture moving company example, the hiring policy classifies all male applicants as hired, and half of female applicants with lifting ability as hired. In this version, the incremental hiring probability for female applicants without lifting power is -45% (below the population hiring average of 95%). Notably, SHAP attributed 17 times as much weight to gender as to lifting ability in its explanation. This version of the example does not satisfy the independence condition. If we restructure the data in the example to satisfy independence, while preserving the marginal distributions, we will indeed get different attributions.

Before describing the results, let me mention one caveat about technique. With independence, it becomes necessary to know how the hiring policy will treat female applicants with lifting ability, as there will be such applicants in the revised pool. For the sake of discussion, assume that the policy prescribes hiring all female applicants with lifting ability.

Imposing independence results in 72.2% of the incremental hiring probability for female applicants without lifting power to be attributable to gender, down from 94.4% in the example without independence. This change reduces the gender-to-lifting ability multiple of 17 for the case non-independence version of the example to 2.6 for the independence version.[37]

For male applicants without lifting ability, imposing independence results in all of the attribution weight being assigned to gender. This is more intuitive than the 150% / -50% split associated with non-independence. It is important to understand that the technical reason the intuitive result obtains with independence is the following: The imposition of independence preserves the population hiring average, and in addition forces the hiring rate for all applicants without lifting ability to be the population average.

For male applicants with lifting ability, independence makes no difference: the result is equal attribution weights.

Being hired is not just about having a hiring probability that is at or above average. Females without lifting ability will want to know what they can do to be hired, in terms of changing their features. In this respect, they are interested in what are technically known as "counterfactual explanations" which are "actionable." [38] The attribution analysis provides them with an indication about the value of improving their lifting ability. If most of the attribution weight is assigned to gender, then the explanation is not actionable in that improving lifting ability (while maintaining gender) will be a waste of time and effort.

A point worth noting is that for the purpose of computing relative attribution weights in the example featuring hiring bias, what we require is that gender and lifting ability be independently distributed. In this regard, the actual probabilities are not as important. Even uniform probabilities are fine. Again, the key issue is that lifting ability be associated with the population average in respect to hiring probability.[39]

The points made in the last paragraph will strike some as counterintuitive; and they are. They are, because of the implication, based on the example, that AI validators would be better off by discarding some distributional information, such as information about feature covariance, and even using inaccurate marginals. However, remember that Mr. WH managed to find workarounds to manage his disability and that these workarounds were successful in most environments. In similar fashion, the explainable AI firms can employ workarounds for SHAP, or more correctly structure environments in which SHAP provides more accurate attributions.

5. Behavioral issues

There are several behavioral issues associated with both AI and explainable AI. Certainly, salience is one, as location of the major flaw in the original SHAP protocol is far from salient. In this section, I discuss additional specific issues.

The team wondered if he would ever fit in.

5.1 Robustness and overconfidence about explainability

Explainability means different things to different people. At a high level, it means having a deep understanding of causal relationships. At a lower level, it means having a sense of how inputs might translate into outputs, but not at a causal level. In this respect, recall that Mr. WH developed workarounds for identifying the people with whom he was interacting, not because he recognized their faces, but because he had identified cues that enabled him to make educated guesses. When healthy, Mr. WH had a higher level of understanding than he did after becoming ill.

Most explainable AI is about making educated guesses. Notably, there are many explainable AI techniques. As discussed above, there are several versions of SHAP. There is an alternative approach to SHAP called "integrated gradients" which is based on averaging partial derivatives in neighborhoods around values of functions being explained.[40] It is one thing to choose an explainable AI model that fits the data best. It is another thing to infer that arriving at a good fit is the same as identifying the underlying process.

For example, in the Apple Card situation discussed above, explainable AI might identify the variable "prime holder" as explaining the differential treatment between males and females. Statistically, that identification might suggest the absence of gender bias. However, if the financial system by default assigns males to be prime account holders, with no sound financial reason, then identifying a role for prime account holder does not imply the absence of gender bias.

Consider a related example involving criminal record and race, where one model assigns most of the attribution weight to criminal record, another model assigns most of the attribution weight to race, and both models have similar predictive power. The lack of robustness stems from the very different attribution weights accorded by the two models. Robustness risk emerges when one model is selected over the other, not just for its predictive power but also for the causal explanation it provides. This issue relates to the Rashomon effect, having a multiplicity of good models which vary widely in their internal structures.

Risk managers need to be aware that there is a risk of becoming overconfident about understanding what a black box algorithm is doing, after discovering a surrogate model that appears to mimic what that black box is doing. Readers seeking a real world example about overconfidence and ML-based decisions can look to the case of the large U.S. real estate firm Zillow. In November 2021, Zillow announced that it was laying off 25% of its workforce, because of a problem with its ML/algorithmic approach to predicting home prices.. Risk managers need to be aware of the importance of having robust explanations, meaning explanations which can be generated in multiple, preferably independent ways, using several explanation methods. Notably, researchers working in AI view SHAP models as lacking in robustness for non-linear black box models, because they only offer local explanations.[42]

5.2. Self-serving bias

As I mentioned previously, an essay by the chief technology officer at ZestFinance was quite critical of SHAP, who wrote on the firm's blog: "There are many details you need to get right in this process, including the appropriate application of sample weights, mapping to score space at the approval cut-off, sampling methods, and accompanying documentation. Out of the box, SHAP doesn't allow you to easily do this." In response to the criticism, Lundberg, who is one of SHAP's academic co-authors, stated that he thought the authors of the critique "have a solid understanding of both finance and explainability and are doing good things. Unfortunately, I think this particular article was influenced too much by 'marketing pressures' that encourage Zest to position itself as the go-to for explainable finance." Lundberg's response, I believe, speaks for itself, in respect to suggesting the presence of self-serving bias.

5.3. Insufficient information sharing

A key finding in the behavioral decision literature is that when people work together in groups and are susceptible to groupthink, they become reluctant to share information.[43] Putting processes in place to facilitate information sharing about ML-based decision making will be especially important for risk management teams that have differential skill sets in respect to data science. Silo structures come about

naturally, for example when a data science team builds a model, communicates it to the risk management team to review, after which a separate ML-operations team takes responsibility for the resulting product. To address the tendency for silo-based information structures, organizations will have to give some thought to developing dashboards that are common to the different teams, providing shared information to help coordinate the teams' various ML-centric efforts.

"We have a really bad case of spotted groupthink."

5.4. Reference points

Recall that SHAP frames attributions using a reference point, the base hiring rate for the population. Reference points are a central feature of prospect theory, the psychological framework developed by Daniel Kahneman and Amos Tversky for explaining how people make choices in the face of risk and uncertainty.[44] Kahneman was awarded an economics Nobel for this work. Prospect theory emphasizes that psychologically, people tend to evaluate outcomes as gains and losses relative to a reference point. Moreover, the psychological values they attach to losses is typically more than twice the value they attach to gains. This issue can become important when users of SHAP interpret its results, in that subgroups with below average hiring probabilities might experience larger psychological effects than those with above average hiring probabilities.

5.5. Simple solutions and complex decision tasks: less is more

The last set of behavioral issues discussed in this section relates to the literature on bounded rationality that grew out of the work of psychologist Robyn Dawes.[45] Below I discuss three aspects of this literature, involving the concepts "less is more," "tallying" otherwise known as the "1/n heuristic," and "fast and frugal trees."

Dawes studied the question of what differentiates experts from non-experts. He concluded that the most important differentiator is experts' ability to identify which among many features are the most important, and their associated directional effects. One of his most striking findings was that simple linear models, based on a small set of features, could outperform the judgments of experts. This work suggested that for many people, too much information would be distracting, and that as a result, these people could improve their judgments by focusing on small amounts of relevant information and ignoring the rest. This perspective came to be called "less is more."

"Less is more" applies to explainable AI in that improvements to the basic SHAP protocol are achieved by assuming that features are independently distributed, thereby discarding information involving feature covariance.

5.6. Simple solutions and complex decision tasks: tallying

Dawes's finding that linear regression models typically outperform expert judgment was surprising to most. However, even more surprising was his finding that using linear models with equal weights instead of regression weights would result in judgments superior to those of experts. The finding about equal weighting is an example of an approach called "tallying" or the "1/n" heuristic. Tallying applies to SHAP when empirical distributional features are replaced by uniform distributions. Notably, the uniform distribution assumption provides more accurate information about relative feature attribution than the empirical distribution.

5.7. Simple solutions and complex decision tasks: pivot points and trees

The issues involving the counterintuitive idea of discarding information are complex. Part of the story relates to what occurs at steps 1, 7, and 10 of the SHAP protocol. However, there are other issues to think about. In developing SHAP, Lundberg and Lee combined information about decision criteria and distributional information. In a sense, the approach conflates two different issues. The first issue pertains to the nature of the decision rule, such as the degree to which it is fair. The second issue pertains to the empirical question of how widespread is the lack of fairness.

In this respect, keep in mind that SHAP is an acronym, standing for "SHapley Additive exPlanations," with the Shapley referring to "the Shapley value." SHAP is not the only way to apply the Shapley value to attribution analysis. A different way involves an examination of the tree structure of the hiring decision.

In the furniture moving company example, think about the hiring manager considering information about candidates' features in a specific order. Notably, the manager might not have to examine information about every feature when deciding whether or not to hire an applicant. In the version of the example illustrating blatant bias, if information is first presented about gender, the hiring manager can make the hiring call without having to examine information about lifting ability.

Here is how the Shapley value would be computed for the hiring decision tree in the blatant bias version of the example. For all applicants, there are two orders in which the information can be presented. The first is gender followed by lifting ability, and the second is lifting ability followed by gender. Consider male applicants with lifting ability. These applicants begin with a score of 0 to denote a status of "no decision." Along the decision tree, at some point the status will convert either to -1 (for rejected) or to 1 (for hired). Once converted, status does not change, no matter what information might flow subsequently.

Begin with male applicants having lifting ability. In both feature presentation orders, hiring status converts from 0 to 1 (hired) only after the information about gender is presented. In this case gender receives a score of 1, and lifting ability receives a score of 0 for both orders. When these are averaged, the Shapley value for gender is 1 and for lifting ability is 0. The analysis for male applicants with no lifting ability produces the same Shapley values. The analysis is the same for female applicants, except that the resulting Shapley value for gender is -1 instead of 1, as all female candidates are rejected.

Focus next on a different hiring policy, in which only males with lifting ability are classified as hired. In this case, information on both features needs to be known before hiring status can convert from 0 to 1.

For male applicants with lifting ability, lifting ability receives a score of 1 when information about lifting ability is presented last, and 0 when it presented first. Gender receives a score of 1 when information about gender is presented last, and 0 when it is presented first. Averaging the scores for gender and lifting ability across orders of presentation produces Shapley values of 0.5 for both gender and lifting ability.

For male applicants without lifting ability, the same analysis applies, except that the final outcome is -1 instead of 1. In this case, information about gender is never pivotal, whereas information about lifting ability is always pivotal. Therefore, the Shapley value is -1 for lifting ability and 0 for gender. For female applicants without lifting ability, the hiring decision is clearly -1 at the first stage, and symmetric for gender and lifting ability, leading to Shapley values of -0.5 for both features. For female applicants with lifting ability, the -1 outcome is only determined when information about gender is presented. Therefore, for these applicants the Shapley value for gender is -1 and for lifting ability is 0.

The bounded rationality approach features a concept known as "fast and frugal trees."[47] This concept involves structuring decision trees to minimize the amount of information required to make a decision. In effect, the Shapley value for a specific feature, when applied to a particular applicant, describes the fraction of trees whereby the feature in question is the pivotal player, in a particular direction. Being a pivotal player is the core concept underlying Shapley's analysis of bargaining power in cooperative games.

For the algorithm in which any and all male applicants are hired, and no other, the Shapley values for gender are 1 for both males with lifting ability and males without lifting ability. The Shapley values for gender associated with female candidates are -1 for both those with lifting ability and those without lifting ability. The Shapely values associated with gender are distinctly different for males and females, which signals the lack of fairness.

Notice that the Shapley value analysis of the decision tree makes no use of distributional information. The attribution analysis is based entirely on the classification algorithm. Now, with attribution analysis in hand, it then makes sense to focus on the empirical (true) distributional information, in order to assess the relative importance of bias in the population to which the algorithm is being applied. In the decision tree-based attribution analysis, algorithmic bias and distributional information are not conflated, as is the case with SHAP. Conflation enters as an issue when computational costs are involved, as conflation can reduce these costs.

Coming back to tallying, the version of SHAP associated with the assumption of uniform distributions comes closest to capturing the tree-based Shapley value analysis, as the latter treats all tree paths equally. For the policy which only classifies male applicants as being hired, the two approaches provide the same information, except that uniform-SHAP frames the information about attribution in terms of incremental hiring probability, whereas tree-based Shapley value presents the information as a decomposition of either 1 or -1.[48]

Robustness is an issue here, given the different Shapley value-based approaches for explaining the nature of the classification algorithm. Each Shapley value-based approach is local in nature and lacking in robustness.

Individually, no single approach reveals the underlying nature of the situation in the "real world." Individually, no single approach is robust. Yet taken together, the different approaches help to provide insight about real world relationships embedded within the black box.

5.8 Confusing prediction tasks with those that identify causation

SHAP is a tool to aid in tasks involving prediction, not necessarily tests to identify causation. This is important because for psychological reasons many people are subject to some form of attribution bias, meaning that they are prone to make specific types of mistakes in the course of drawing inferences about what causes what. In particular, people are prone to confuse the notions of predictability and causation.[49]

To illustrate attribution bias, consider the moving company example discussed above. Suppose that this company offers one year contracts to several types of commercial customers, such as real estate firms who stage the homes they list on the market. Some customers have a stronger need for the moving company's services than others. Imagine that the company manages its spare capacity by offering lowering contract prices to its marginal customers, meaning customers whose needs for the company's moving services is marginal. Imagine too that the moving company monitors the quality of its service by relying on customer feedback, especially complaints. When providing accurate feedback is costly, customers with strong needs are more likely to provide feedback generally, especially in respect to complaints.

Suppose that the likelihood of contract renewal depends on the strength of a customer's need for the moving company's services, the price, and the quality of the service. In this situation, it is plausible that customers which pay a higher price and complain more, because of their strong need for the moving company's services, will be more likely to renew their contracts than other customers.

Consider a prediction task. Suppose that the moving company wants to use a SHAP approach to predict which of its customers are most likely to renew their contracts, based on features such as price paid and frequency of complaints. Then it might make sense for them to predict that customers who pay higher prices and communicate more complaints will be more likely to renew.

Consider next an expected profit maximizing decision task, using price and quality of service as decision variables. Expected profit depends on customer renewal. In line with the correlations just described, would it make sense for the moving company to charge its customers higher prices and offer inferior quality in order to increase the average renewal rate? Probably not, as raising price or reducing quality for a specific customer, with an identified need, will almost surely make the service less attractive.

In a situation such as this one, misapplying a SHAP-based approach, which is not an inherently causal model, to a decision task constitutes an example of attribution bias. That said, there are techniques, some which incorporate SHAP, which can be used to analyze causality, at least in some situations. One such technique is called double/debiased machine learning (double ML), and it is designed to return causal effects.[50] However, double ML is not a panacea, and sometimes it will be important to engage in experimentation in order to identify the impacts of key driving variables which generate confounding effects, such as inherent need for service in the moving company example.

6. Discussion and conclusion

As the use of machine learning decision processes grows, it is becoming increasingly important to monitor ML algorithms, apply explainable technologies to assess what is inside their black boxes, and address issues related to fairness or a lack thereof.

The concept of CDD discussed earlier was introduced in an article by Wachter, Mittelstadt, and Russell (2021). This article places the issue of algorithm-generated discrimination into the broader social context. The authors of this article argue that "intuition can no longer be relied upon, as it has been historically, as the primary mechanism to identify and assess potentially discriminatory actions in society. Algorithmic systems render a fundamental mechanism of European Union non-discrimination law useless, necessitating new detection methods and evidential requirements." New detection methods involve explainable AI.

At the same time, Wachter et al. argue that "fairness cannot and should not be automated." Instead they propose a set of summary statistics, which they call conditional demographic disparity (CDD).[51] They suggest that CDD can serve as "a baseline for evidence to ensure a consistent procedure for assessment (but not interpretation) across cases involving potential discrimination caused by automated systems."

CDD is important, as it helps identify situations that on the surface suggest the presence of bias, even though no bias is present. For example, the university admissions example described in the introduction is one that suggests the presence of bias on the surface, even though no bias is present.

CDD and explainability techniques are complementary. Both provide answers to important, but generally different, questions about bias. CDD is especially important for identifying the presence of bias in legal proceedings. Explainability provides insight into the relative weights of various features in determining how classification is accomplished.

There is an argument to be made for the human touch, what Wachter et al. call "humans in the loop." Most husbands do not mistake their wives for hats, as did Mr. WH. In some circumstances, a well-functioning human touch might help mitigate the problems that resulted from Mr. WH's deficiency. The same is true for algorithmic decision making, including explainable AI. The argument involves how much of a human touch is required, depending on circumstances, but not whether it makes sense to use explainable AI.

This paper focuses on subtleties associated with the black box character of machine learning algorithms and techniques to infer the nature of what is going inside those black boxes. This will be an important function for risk managers going forward. Explainable AI is valuable, but like ML-based algorithms, can be imperfect. Risk managers will need to have a good understanding of how explainable AI works, and to understand where explainable AI might lead its users astray. The goal of this paper is to use a set of stripped down examples to help risk managers in this regard. A general understanding will be insufficient when intuition cannot be relied upon. The details are important. For risk managers, forewarned precedes becoming forearmed.

7. References

1. http://www3.weforum.org/docs/WEF_AI_in_Financial_Services_Survey.pdf. See also https://blog.fiddler.ai/2020/11/rise-of-banking-ai-validator/

2. https://news.ihsmarkit.com/prviewer/release_only/slug/technology-global-business-value-artificial-intelligence-banking-reach-300-billion-203

3. Regulatory bodies such as the Federal Reserve and Office of the Comptroller of the Currency (OCC) have established guidelines for conducting MRM. For example, see https://www.federalreserve.gov/boarddocs/srletters/2011/sr1107a1.pdf. These guidelines provides standards for activities such as the documentation of model data and model validation.

4. An important function of AI validators is to probe models in order to assess their performance across domains of various granularity, such as individual data instances (local behavior), groups of instances (region behavior) and all data instances (global behavior). AI validators analyze ML-based models in order to assess model robustness (or weakness), an exercise that is especially important given the black box nature of these models, prior to and during the operation of these models.

5. Many ML-generated algorithms are used for decision making in industries which are overseen by regulators governed by legislation such as the Equal Credit Opportunity Act (ECOA), the Fair Housing Act (FHA), the Civil Rights Act, and the Immigration Reform Act.

6. See Hancox-Li, Leif, 2020. “Robustness in Machine Learning Explanations: Does It Matter?” FAT* ’20, January 27–30, 2020, Barcelona, Spain. http://philsci-archive.pitt.edu/16734/1/preprint.pdf.

7. ML-generated algorithms are essentially functions which map inputs, known as features, into outputs. The nature of these function can be difficult to understand from the outside, which is to say that ML-generated algorithms are like black boxes. Lundberg and Lee designed SHAP to analyze the relative contributions of different features when determining how inputs are mapped into outputs. SHAP is very versatile and is used to analyze text as well as numerical data, once text is transformed so that it comes to be represented as a set of numerical features.

8. For example, both Amazon and Fiddler AI offer monitoring products.

9. See https://www.youtube.com/watch?v=BYjLW9wle7k for a discussion between Sri Krishnamurthy, founder of QuantUniversity and Krishna Gade, founder of Fiddler AI. Krishnamurthy made the following statement: “If you are a company which gives millions and millions of credit cards to people, and you have built a model which has either an explicit or an implicit bias in it, there are systemic effects.” A Fiddler AI blogpost commented on this statement, remarking: “That’s why it’s so important for the financial industry to really understand models and have the right practices in place, rather than just rushing to adopt the latest AutoML or deep neural network technologies.”
See https://blog.fiddler.ai/2021/05/achieving-responsible-ai-in-finance-with-model-performance-management/.

10. ML-based technology need not feature bias, and bias can certainly emerge from decision systems that do not rely on ML.

11. Not all ML-algorithmic errors feature bias.

12. See Wachter, Sandra, Brent Mittelstadt, and Chris Russell, 2021. “Why Fairness Cannot be Automated: Bridging the Gap Between EU Non-discrimination Law and AI.” Working paper. There is controversy about whether the case involving UC Berkeley is apocryphal. However, it certainly illustrates a statistical issue known as “Simpson’s Paradox.”

13. See Metz, Cade, 2021. “Using A.I. to Find Bias in A.I.,” The New York Times, June 30. https://www.nytimes.com/2021/06/30/technology/artificial-intelligence-bias.html. This article discusses several firms that are working on explainability issues. Most are small. Large firms, such as Amazon which offers a product called SageMaker, and Google which offers a product called xAI, are active on this front. See Sarkar, Tirthajyoti, 2019. “Google’s new ‘Explainable AI” (xAI) service.”
https://towardsdatascience.com/googles-new-explainable-ai-xai-service-83a7bc823773.

14. See “Perspectives on Artificial Intelligence: Where We Are and the Next Frontier in Financial Services.” https://financialservices.house.gov/calendar/eventsingle.aspx?EventID=403824

15. https://www.nytimes.com/2021/04/16/business/artificial-intelligence-regulation.html. See also Wachter, Sandra, Brent Mittelstadt, and Chris Russell, 2021. “Why Fairness Cannot be Automated: Bridging the Gap Between EU Non-discrimination Law and AI.” Working paper. In this paper, the authors argue that “automating fairness or non-discrimination in Europe may be impossible because the law, by design, does not provide a static or homogenous framework suited to testing for discrimination in AI systems… Compared to traditional forms of discrimination, automated discrimination is more abstract and unintuitive, subtle, intangible, and difficult to detect.”

16. https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html and https://www.dfs.ny.gov/reports_and_publications/press_releases/pr202103231.

    The press release from the New York State Department of Financial Services, pertaining to the outcome of its investigation of Apple Card and Goldman Sachs stated the following: “In terms of gender, the Department found, based on its data analysis, that Apple Card applications from women and men with similar credit characteristics generally had similar outcomes. For all consumers who reported concerns about their Apple Card credit application outcomes to the Department, evidence showed that those decisions were explainable, lawful, and consistent with the Bank’s credit policy. However, the Department concluded, deficiencies in customer service and a perceived lack of transparency undermined consumer trust in fair credit decisions.

    The report notes that Goldman Sachs and Apple have since taken steps to improve transparency, implemented a program to assist denied applicants in improving their credit with the goal of obtaining Apple Card approval, and modified a policy that previously required approved applicants to wait six months before appealing credit terms.” The press release also stated: “Consumers voiced the belief that if they shared credit cards with spouses, even if only as authorized users, they were entitled to the same credit terms as spouses. In reality, however, underwriters are not required to treat authorized users the same as account holders and may consider many other factors.”

17. https://www.nytimes.com/2021/04/16/business/artificial-intelligence-regulation.html

18. For example, explainable AI can be used to structure questions of an ML-generated algorithm for evaluation credit applicants. An AI validator using explainable AI might ask questions such as the following: Looking both globally and across groups, what were the main features that impacted approval or rejection outcomes? Globally or across specific groups? For individual loan applicants, which specific features appeared to be most responsible for decisions about approval or rejection? Are there any features which appear to stand out for some individual applicants relative to others in their group? What kinds of changes might an applicant make in order that a decision involving rejection be changed to one for approval?

19. At the time, Douglas Merrill was CEO at ZestFinance. He had founded the firm in 2009.

20. The critique on SHAP mentioned in this testimony appeared as a post entitled “Why Lenders Shouldn’t ‘Just Use SHAP’ To Explain Machine Learning Credit Models,” Posted on May 30, 2019 by Jay Budzik. https://www.zest.ai/insights/why-lenders-shouldnt-just-use-shap-to-explain-machine-learning-credit-models

21. See Lundberg, Scott and Su-In Lee, 2017. “A Unified Approach to Interpreting Model Predictions,” 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. https://arxiv.org/pdf/1705.07874.pdf.

22. For a discussion of robustness, see Hancox-Li, Leif, 2020. “Robustness in Machine Learning Explanations: Does It Matter?” FAT* ’20, January 27–30, 2020, Barcelona, Spain http://philsci-archive.pitt.edu/16734/1/preprint.pdf. Hancox-Li uses the terminology of surrogate functions and distinguishes between local explanations, global explanations, and counterfactual-based explanations.

23. SHAP has come to be used by practitioners in financial services, especially to improve processes related to lending and credit management. Some SHAP users discovered that the technique had flaws, and in consequence developed alternative approaches. As was pointed out above, in 2019, during Congressional testimony about the growing role of AI in finance, mention was made of SHAP and its associated flaws. Through 2019 and 2020, data scientists on these issues, some by making relatively modest modifications of the original SHAP approach.

24. Merrick, Luke and Ankur Taly, 2020. “The Explanation Game: Explaining Machine Learning Models Using Shapley Values,” Part of the

Lecture Notes in Computer Science book series (LNCS, volume 12279). The authors were with the firm Fiddler AI at the time they wrote this paper.

25. This type of policy is also analyzed in Merrick and Taly (2020).

26. Of course in an example as simple as the one used here, ML can be viewed as overkill. The point, of course, is to use a simple example to explain how explainable AI analyzes a policy, and what ML generates are policies, mappings from inputs to classification outputs.

27. This example shares some of the aspects of the UC Berkeley admissions situation discussed in connection with Simpson's Paradox. In the moving company example, information about type of specific task, physical or logistical, corresponds to a specific academic department at UC Berkeley. There are some situations where the analysis is conducted, despite information on important features not being available.

28. See Wachter, Sandra, Brent Mittelstadt, and Chris Russell, 2021. "Why Fairness Cannot be Automated: Bridging the Gap Between EU Non-discrimination Law and AI." Working paper.

29. For an informal discussion of the development of CDD, see https://www.amazon.science/latest-news/how-a-paper-by-three-oxford-academics-influenced-aws-bias-and-explainability-software, from which the following excerpt is taken:
CDD is defined as "the weighted average of demographic disparities for each of the subgroups, with each subgroup disparity weighted in proportion to the number of observations it contains."

    "Demographic disparity asks: 'Is the disadvantaged class a bigger proportion of the rejected outcomes than the proportion of accepted outcomes for the same class?'" explained Sanjiv Das, the William and Janice Terry professor of finance and data science at Santa Clara University's Leavey School of Business, and an Amazon Scholar.

30. The name Simpson's Paradox derives from the work of Edward H. Simpson, who in 1951 described this phenomenon. See Simpson, Edward H. (1951). "The Interpretation of Interaction in Contingency Tables," Journal of the Royal Statistical Society, Series B. 13: 238–241.

31. For a discussion of computational issues associated with SHAP, see https://christophm.github.io/interpretable-ml-book/shap.html#disadvantages-15.

32. In this example, gender, not lifting ability, is understood to define the protected class, meaning the class of people for whom there is a social concern about being discriminated against. In many applications, gender would not be an actual input variable for an ML-generated algorithm, although in this example, it is.

33. For the loan application example, the general acceptance rate is 70%. For high net worth male applicants, SHAP decomposes the 30% incremental probability into equal increments. For low net worth male applicants, SHAP decomposes the -70% incremental probability into 9% for gender and -79% for net worth. For high net worth female applicants, SHAP decomposes the -70% incremental probability into -79% for gender and 9% for net worth. For low net worth female applicants, SHAP decomposes the -70% incremental probability into equal increments.

34. Implementing SHAP in this case requires that we know the hiring policy for female applicants with lifting ability, something not required when the actual probabilities are used. When no female applicants with lifting ability are present in the population, we cannot be certain that a hiring policy is fair. The most we can say is that we have no evidence that it is unfair. If we apply SHAP to the version of the example in section 2 which appears to be fair, we will find that SHAP does not attribute 100% of the incremental probability weight to lifting ability. Only if we invoke independence, and set the hiring policy so that female applicants with lifting ability are hired, the same as male applicants, will SHAP attribute 100% of the incremental probability weight to lifting ability.

35. For the loan application example, applying SHAP with independence also produces decompositions of incremental probabilities that are more in accord with intuition than SHAP applied to the actual distribution. However, invoking the independence modification only impacts the decompositions for the case of low net worth male applicants and high net worth female applicants. For low net worth male applicants, SHAP decomposes the 70% incremental probability into 5% for gender and -75% for net worth. For high net worth

female applicants, SHAP decomposes the -70% incremental probability into -75% for gender and 5% for net worth. SHAP treats male applicants with low net worth and female applicants with high net worth as mirror images of each other, and does not indicate whether any alleged discrimination stems from gender or net worth. If female applicants are viewed as a protected class, as opposed to applicants with low net worth, then this extra information leads to a focus on the decomposition for female applicants having high net worth, and therefore to the overweight (-75% relative to +70%) associated with gender for these applicants.

36. In the example without independence, consider the case of male applicants with lifting ability. Here, SHAP splits the attribution between gender and lifting ability equally, a mistake which occurs for the following reason: When SHAP encounters a male applicant with lifting ability, and assesses incremental probability associated with lifting ability (before doing so for gender), it mistakenly assigns the incremental probability to lifting ability instead of to gender. SHAP gets fooled at this point. It mistakes lifting ability for gender, just as Mr. WH mistook his wife for a hat.

37. Another variation of blatant bias is when the hiring algorithm classifies all male applicants as hired, half of female applicants without lifting ability as hired, and no female applicants with lifting ability as hired. There is a sexist interpretation to this version of the example. The interpretation combines a male hiring manager with misogynist views, along with relative physical attractiveness of female applicants being negatively correlated with lifting ability. Here, I assume that physical attractiveness, while observed by the male hiring manager, is an omitted variable in respect to the inputs being measured by outside observers. With independence imposed, for female applicants with lifting ability, SHAP attributes 84.2% of the incremental hiring probability to gender. For female applicants without lifting ability, the corresponding figure is 127.8%, indicating that the reason for these applicants being classified as hired has everything to do with gender. For males with lifting ability, gender receives 150% of the attribution, again strongly pointing to gender discrimination. This version of the example focuses on a human hiring manager, with the ML-aspect coming in during the training phase, based on data involving past human decisions.

38. See Galhotra, Sainyam, Romila Pradhan, and Babak Salimi, 2021. “Explaining Black-Box Algorithms Using Probabilistic Contrastive Counterfactuals.” Working paper. University of Massachusetts, Amherst. https://arxiv.org/pdf/2103.11972.pdf.

39. Using uniform probabilities alters the average hiring probability, and therefore the resulting values produced by SHAP. In the data, the percentage of the hiring pool that are males with lifting ability is 50%, but under the uniform distribution it is 25%. Consider the blatant bias version of the example in which the hiring policy classifies all males as hired and all females as rejected. For male applicants with lifting ability, gender and lifting ability receive attribution probabilities of 37.5%, which sum to 75%. For male applicants without lifting ability, gender receives 12.5% while lifting ability receives -37.5%. For female candidates with lifting ability, gender receives -37.5% while lifting ability receives 12.5%. For female candidates without lifting ability, gender receives -12.5% and lifting ability receives 12.5%.

40. See Shanbhag, Aalok Avijit Ghosh, and Josh Rubin, 2021. “Unified Shapley Framework to Explain Prediction Drift,” https://researchain.net/archives/pdf/Unified-Shapley-Framework-To-Explain-Prediction-Drift-1933830

41. See Dong, Jiayun and Cynthia Rudin. 2019. “Variable Importance Clouds: A Way to Explore Variable Importance for the Set of Good Models,” arXiv e-prints, Article arXiv:1901.03209 (Jan 2019). arXiv:stat.ML/1901.03209.

42. In respect to Zillow,see Picchi, Aimee, 2021. “Zillow to Lay Off 25% of its Workforce and Shutter House-flipping Service.” https://www.cbsnews.com/news/zillow-layoffs-closing-zillow-offers-selling-homes/ In respect to local explanations, see Alvarez-Melis, David and Tommi S. Jaakkola. 2018. “On the Robustness of Interpretability Methods.” http://arxiv.org/abs/1606.03490; Alvarez-Melis, David and Tommi S Jaakkola. 2018. “Towards Robust Interpretability with Self-explaining Neural Networks.” In Proceedings of the 32nd International Conference on Neural Information Processing Systems. 7786–7795.

43. For a discussion of this issue, see Shefrin, Hersh, 2016. Behavioral Risk Management: Managing the Psychology That Drives Decisions and Influences Operational Risk, New York Palgrave Macmillan.

44. See Kahneman, Daniel and Amos Tversky, 1979. “Prospect Theory: An Analysis of Decision under Risk,” Econometrica, 47(2), 263-291.

45. Dawes, R. M., 1979. “The robust Beauty of Improper Linear Models in Decision Making,” American Psychologist, 34(7), 571–582.

46. See the work by two Google data scientists: Sundararajan, Mukund and Amir Najmi, 2020. “The Many Shapley Values for Model Explanation,” Proceedings of the 37th International Conference on Machine Learning, PMLR 119:9269-9278. http://proceedings.

mlr.press/v119/sundararajan20b.html. In addition, for a discussion of the application of SHAP for tree-based problems, see https://christophm.github.io/interpretable-ml-book/shap.html#treeshap.

47. For a discussion of "less is more," "tally," and "fast and frugal trees," see Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In G. Gigerenzer, P. M. Todd, & The ABC Research Group, Simple Heuristics that Make Us Smart (pp. 3–34). Oxford University Press. See also Martignon and U. Hoffrage, 1999. "Why Does One-Reason Decision Making Work? A Case Study in Ecological Rationality," In G. Gigerenzer, P. M. Todd, & The ABC Research Group, Simple Heuristics that Make Us Smart (pp. 119–140). Oxford University Press.

48. Consider uniform-SHAP when the algorithm prescribes only hiring males with lifting ability. In this case, the percentage of the hiring pool that are males with lifting ability is 25%. As was mentioned earlier, for male applicants with lifting ability, gender and lifting ability receive attribution probabilities of 37.5%, which sum to 75%. For male applicants without lifting ability, gender receives 12.5% while lifting ability receives -37.5%. For female candidates with lifting ability, gender receives -37.5% while lifting ability receives 12.5%. For female candidates without lifting ability, gender receives -12.5% and lifting ability receives 12.5%. Consider tree-based Shapley. For male applicants with lifting ability, the Shapley values will be 0.5 for both gender and lifting ability. For male applicants without lifting ability, presenting information on gender first maintains the hiring status at 0. Presenting information on gender second maintain the hiring status at -1. Therefore, gender receives a Shapley value of -0.5. For lifting ability, presenting lifting ability second changes the hiring status from 0 to -1. Presenting lifting ability first also changes the hiring status from 0 to -1. Therefore, for male applicants without lifting ability, lifting ability receives a Shapley value of -1, twice the value for gender, because lifting ability is pivotal twice as often as gender is pivotal. Notice that the relative attribution weights are different in this case, (gender to lifting ability) for the two approaches, uniform-SHAP and tree-based Shapley.

49. See the joint article about causality and interpretable machine learning, entitled "Be Careful When Interpreting Predictive Models in Search of Causal Insights" by Eleanor Dillon, Jacob LaRiviere, Scott Lundberg, Jonathan Roth, and Vasilis Syrgkanis: https://shap.readthedocs.io/en/latest/example_notebooks/overviews/Be%20careful%20when%20interpreting%20predictive%20models%20in%20search%20of%20causal%C2%A0insights.html. See also Janzing , Dominik, Lenon Minorics, and Patrick Blobaum, 2019.b "Feature Relevance Quantification in Explainable AI: A Causal Problem,"Amazon Research Tubingen, Germany.

50. Double ML searches for the presence of confounding factors associated with causal relationships. Suppose that a risk manager investigates whether a variable X serves as a cause for variable Y. The risk manager might be concerned that although X and Y are statistically related, that relationship might stem from the presence of a known confounding or an unknown confounding factor. To investigate this issue, the risk manager can run two ML routines, in other words, a double ML routine, one pertaining to the relationship between X and the known confounding factors, and a second ML routine pertaining to the relationship between Y and the known confounding factors. To identify whether there might be an unknown confounding factor, the risk manager examines the residuals from both routines to see whether they are statistically related. Such a relationship would signal the possibility of a common confounding factor.

51. To illustrate DCC, consider the example discussed earlier featuring a token 1% of female applicants with lifting ability, so that 9% of female applicants have no lifting ability, 50% of male applicants have lifting ability, and 40% of male applicants have no lifting ability. Focus on the case in which the hiring algorithm recommends hiring all males and anyone with lifting ability.

    Begin by comparing the percentage of hired applicants who are female to the percentage of rejected applicants who are female. The former is 1.1% (=1% / (1% + 50% + 40%)), while the latter is undefined 100% (= 9%/9%). Notice the string of inequalities 100% > 50% > 1.1%. This string signals a condition known as "demographic disparity with negative dominance," and it suggests the presence of a hiring algorithm which is biased toward female applicants.

    Next, consider the same comparisons but for two subgroups: those with lifting ability and those without. For those with lifting ability the string is "undefined (take it as infinity) > 50% > 2%." For those without lifting ability, the string is "100% > 50% > 0%." Remember that for those without lifting ability, all male applicants are hired but no female applicants are hired, which is what the 0% signifies. If we agree that both inequality strings hold, then we say that this situation features CDD, conditional demographic disparity.

Professional Risk Managers' International Association

PRMIA
INSTITUTE
Professional Risk Managers'
International Association