National Institute of Standards and Technology
100 Bureau Drive, Stop 2000
Gaithersburg, MD 20899

RE: RFI: Developing a Federal AI Standards Engagement Plan

As a leading global professional services company, Accenture provides a broad range of services and solutions in strategy, consulting, digital, technology and operations that span multiple industries. We combine artificial intelligence (AI) with deep industry and analytics expertise to help our clients embrace these emerging, intelligent technologies confidently and responsibly.

Accenture is grateful for the opportunity to provide input to the National Institute of Standards and Technology (NIST) Federal AI Standards Engagement Plan. We applaud the Administration in its commitment to ensure the U.S. maintains global leadership in AI, while driving efforts to advance trust and transparency so the American people may realize the positive benefits of these technologies.

Accenture is partnering with business leaders across industries who want to leverage AI to grow, increase productivity, improve efficiencies, and innovate new solutions. As part of those strategies, we work to ensure value and technology are integrated at every step. We advise clients to take a people-first approach to AI; to incorporate data privacy, safety, and cybersecurity; and to participate in multi stakeholder opportunities, such as the NIST process, which can help ensure producer and consumer interests while not hindering innovation or the enterprise economy.

Accenture believes that a strong AI internal governance framework is crucial for any organization developing and/or deploying AI. An organization's AI governance framework should include ongoing assessments and rigor around responsible and ethical design and use. As AI quickly becomes a major tool for both customer and citizen services, it will be essential to promote trust and transparency by ensuring clear governance is in place to establish boundaries on what AI systems can and will be used for, how fairness will be measured, what data will power them, how users will interact with them, and how adjustments will be made, when appropriate. With this submission we hope to address the following questions posed by this RFI:

> NIST question: AI technical standards and tools that have been developed, and the developing organization, including the aspects of AI these standards and tools address, and whether they address sector-specific needs or are cross-sector in nature.
>
> *Accenture: The market is **already adopting standardized AI** to grow, increase productivity, improve efficiencies, and innovate new solutions. While AI that operates on the edge (or real-time AI machine learning) exists and those cases continue to grow, the current state of AI is such that the vast majority of major companies are cautiously proceeding forward with this emerging technology. Now is the right time to engage industry and co-create frameworks for best practices as adoption of AI matures.*
>
> *NIST question:* Whether the need for AI technical standards and related tools is being met in a timely way by organizations.

*Accenture: As industry adoption of AI matures, so do the **safety and performance features**. Accenture continues to drive innovation in both ethical governance frameworks and combating adversarial AI. These are just two examples of how companies are pro-actively seeking answers to AI's biggest challenges.*

NIST question: Technical standards and guidance that are needed to establish and advance trustworthy aspects (e.g., accuracy, transparency, security, privacy, and robustness) of AI technologies

*Accenture: Governance of algorithms must include both **quantitative and qualitative measures** (technical and non-technical). Quantitative measures are the empirical evidence necessary to prove AI systems are effective, fair, and transparent. Qualitative measures enable the critical thinking necessary to interpret evidence effectively. Both context and evidence are necessary to understand risks because there is no one definition of fairness, nor one understanding of sufficient transparency.*

We hope the NIST federal AI standards engagement will lay a foundation that we together can continue to iterate from. We are encouraged that NIST will continue to engage a broad stakeholder audience in this process to ensure that the federal government gathers perspectives beyond data scientists, AI developers and non-IT government program experts. Professionals from across disciplines and interests must work closely together to systematically tackle the opportunities and challenges of AI.

NIST Question #1 - The Current State of AI Adoption: Standardized AI

The Accenture Applied Intelligence Platform

Companies are eager to take advantage of the benefits of AI. In a 2017 survey of 5,400 business and IT executives across 31 countries, more than one-third indicated they were set to make extensive investments in each of seven critical AI technologies.

To help companies take the next steps, Accenture created two indexes to see what has been working so far. We studied both the Fortune Global 100 and what we call the Intelligent Global 100—pioneers in the development of AI technologies and applications—for the period 2010 to 2016. For those 200 companies, we reviewed both their in-house focus (for invention) and their external orientation (for collaboration). Both are essential. Companies will need in-house talent and sometimes proprietary capabilities for AI, as they will need to own some of the technology and some of the data. They will also need to be deeply involved in a broader ecosystem. Neither startups nor incumbents will thrive with a "not invented here" approach. And yet our analysis revealed that fewer than 20 percent score well on both indexes — companies we call "collaborative inventors" — while 56 percent were weak on both.

When Accenture launched our first analytics as a service platform in 2015, AI adoption was just beginning to be adopted by large companies across industries. Since then, we have grown what we now call the *Applied Intelligence Platform (AIP)* to help our clients adopt reliable AI systems at scale and to transform the enterprise through AI. AIP allows organizations – across sectors and industries - to apply pre-configured, and configurable to use case, self-learning industry solutions, and to develop new solutions, without the need for deep data science expertise — which is becoming an increasingly scarce resource. It integrates these capabilities with edge analytics and Internet of Things (IoT) services, as well as provides access to more than 350 curated data sources — all made accessible via an on-demand, low-code software studio. AIP allows our clients to take advantage of the expertise of Accenture's more than 3,000 data scientists and 6,000 deep AI experts, and 1,500 AI related patents As a result, clients are less dependent on specific technologies as the platform leverages solutions and

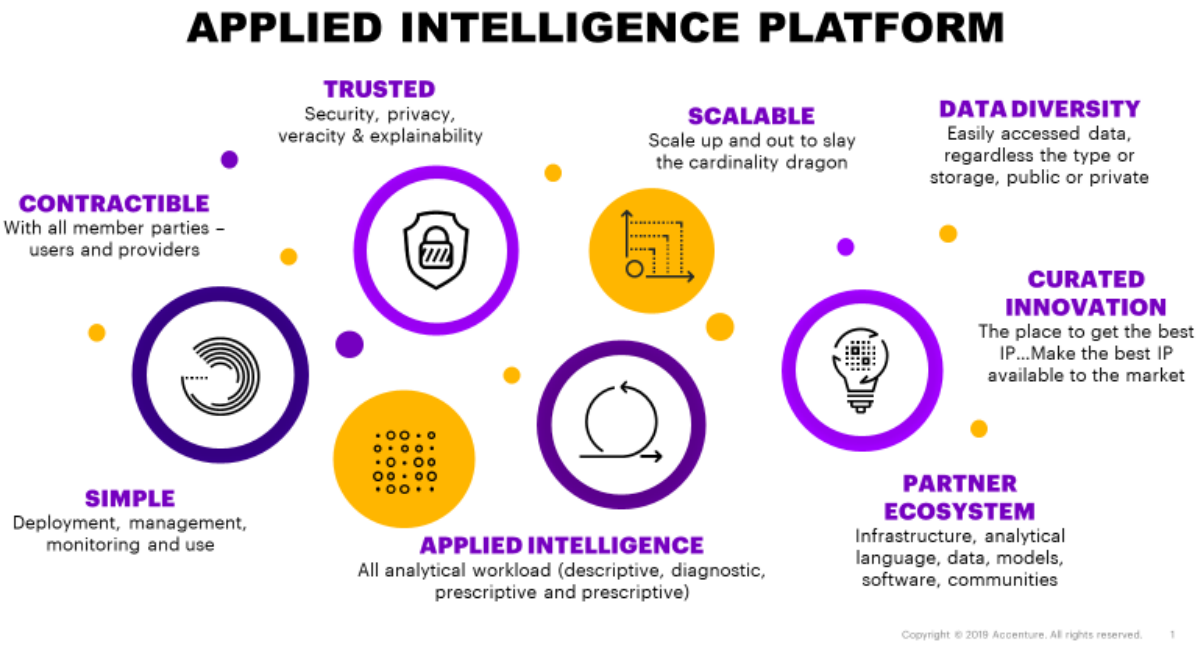
tools from leading technology providers to enable creation of reliable, safe, trustworthy solutions across industries and functions.

Accenture continues to pioneer and integrate the latest and best-in-class applications to ensure that, in an increasingly complex and fast-moving world, we can help our clients as they face unprecedented challenges in: cybersecurity, privacy, safety, data veracity, and preventing bias and discrimination. With many of our technology partners at SAP, Workday, Oracle, and Microsoft, we offer nearly 250 applications on AIP. Accenture has already developed a range of use cases and more than 40 intelligent industry solutions to help bridge the gap between information technology and operational technology. Use cases include claims fraud, asset and fleet management solutions, and energy consumption.

In the face of rapid change, the AIP ecosystem fosters a range of intelligent applications that run models on continuously updated enterprise and device data, which can drive network effects and new business models. We partner with our cloud ecosystem partners, including Amazon Web Services, Azure, and Google Cloud working side-by-side with Accenture Security to harden and secure the AIP and achieve certification with key third-party standards. Robust user identity and access control, including FIPS 140-2 validated cryptography and support for multifactor authentication, provide both safeguards and auditability. Penetration and other vulnerability testing are used to protect against external attacks at the application, software and hardware levels and to validate network isolation. Security design patterns are used for compliance with data privacy laws, such as HIPAA and other regulations governing the use of personal identifiable information, with HiTRUST certification and FedRamp approval.

Accenture creates policies and processes that advance responsible development, deployment and use of AI and does so in a way that benefits all impacted parties and ensures innovation isn't hampered. At the same time, we seek to empower and train our clients' talent to work alongside these technologies to make faster, more informed decisions that are less transactional and more strategic. Our platform has the benefit of being operable across industries, with the ability to adopt the latest compliance requirements quickly and at scale. Currently, AIP, does not operate on the edge. Edge computing requires a significant amount of data science to vet real-time data which is a barrier for many organizations in adopting online AI/machine learning (ML). In the future, custom and edge AI will grow in use, but we predict that time will only come when companies feel they have the expertise and compliance guidance in which to do so.

With each generation of AI, we work with compliance, data scientists, engineers and clients to continue to iterate new tools that can be added to our platform to ensure our clients have the ability to deploy trustworthy, reliable, secure, and ethical AI as soon as the solution is available. In this paper, we will describe two offerings that illustrate how Accenture continues to drive solutions that can be delivered quickly and scale quickly in order to respond to growing challenges: Responsible AI and adversarial AI.



NIST Question #6 – Ensuring the Safety and Performance of AI

Management of the AI Lifecycle

Accenture is increasingly helping our clients to deploy AI. At the same time, we work with our clients to incorporate best practices in management of those systems.

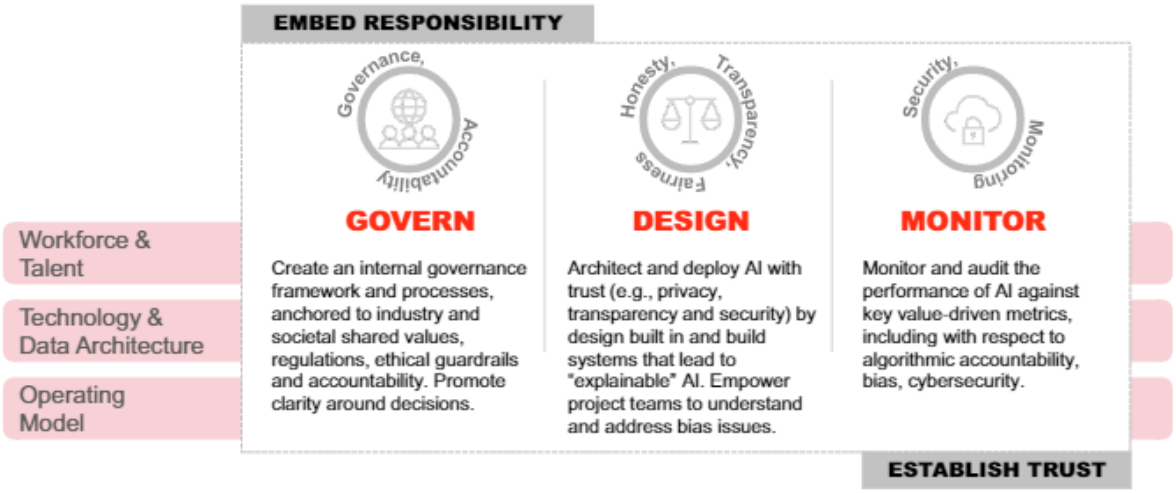
We encourage all organizations to consider that there are two sides to the metaphorical technology coin: both the technology must be properly vetted and trained as well as the operator of said technology. We must take care to consider both sides of the equation.

The first area where organizations often start is *transparency*. Transparency is the ability to understand how and why an AI system decides and acts, particularly in the context of increasingly complex models. Transparency should include two important factors: understandability and interpretability.

> *Understandability* enables a non-technical person (e.g. business executive or customer) to gain insight into how an algorithm works, and why it made a given decision. It is critical that non-technical persons understand how their data is being used and how their actions can generate new predictions. There is an important difference between merely meeting any legal requirements to be transparent versus a desire to establish trust and prioritize understandability.
>
> *Interpretability* allows a technical expert, such as an AI/machine learning expert to understand why an algorithm made a given decision. Interpretability would allow government to know how their models will act in the "real world." Interpretability tends to be the focus of what organizations such as DARPA call "explainable AI". DARPA defines explainability as the ability of machines to: 1) explain their rationale 2) characterize the strengths and weaknesses of their decision-making process and 3) convey a sense of how they will behave in the future.

In addition, organizations may want to consider how to proactively justify their design choices by explaining:

- why they chose a particular data set to draw inferences;
- why these inferences are relevant (and ethical) for the chosen decision they are trying to make; and
- whether the data and methods used to draw the inferences are accurate and statistically reliable for the population they are trying to serve. (A data set full of Iowans would not serve the population of New York whose population is full of different characteristics.)



Human participation is critical to creating AI systems. In considering the human impact of integrating AI into high risk areas, such as criminal justice and health care, organizations can set out to design, build and deploy AI systems whereby human responsibility is enhanced. Whereas transparency provides insights into the systems driving decision-making, human participation enables the ability to change or alter how consumers interact with that system. An added complexity, that is aligned with the issue of transparency, is accountability of human and algorithmic systems. Presumably we build algorithms, at least in part, to standardize and address human bias even while we simultaneously say AI can be biased and call for human oversight. Because of this, we need to remain focused on tracing decision making, not just of algorithms but of people.

In addition, ethics are critical to informing an organization's strategy for its technology deployments. Organizations should consider what are the values that should be encompassed in their product, and how these values might vary across different demographics. Organizations can then proceed by developing AI that incorporates those values.

Policymakers have long used the word "transparency" to address issues in data privacy and security, not algorithmic harms or disparate impact. This is an important distinction, as any discussion of potential harm to impacted communities must consider systemic and institutionalized bias and discrimination as well as systems of power. This means that representative data sets must be selected carefully and that even if an organization is able to build what they consider to be a representative data set, especially in cases that would impact a human life such as criminal justice, health care, and finances, we need to consider implementing systems that enable agency – the ability to take meaningful action against harm. While transparency is necessary, it is insufficient.

Governing bodies should consider establishing flexible guiding principles to govern AI that are general enough to evolve with a rapidly changing technological environment but are also specific enough to be useful for applications. Effective guiding principles marry both restrictive rules and open-ended principles to provide both

the nuance and flexibility required to govern this rapidly developing technology. Furthermore, clearly articulated and established guiding principles create a culture within an organization that allows for ethically responsible attitudes and behaviors from top to bottom.

Ethics Committees

As previously stated, AI ethics should be derived from an organization's core values and mission statement. Rather than create a committee or board that is introduced at the end of product development, a better purpose would be to utilize this resource as an early-stage education tool. No one committee or board is able to encompass all of the skills necessary to provide this nuanced education on all technological implications. A better plan is to use this body as a sourcing group to identify experts from a wide range of backgrounds and empower them to shape design and development.

At Accenture, we are constantly refining our Responsible Business practice, which is an interdisciplinary leadership and practitioner community that encompasses technology, sustainability, legal, corporate social responsibility, and others. We evangelize and apply the principles developed by this group via our Responsible Innovation groups, which includes Responsible AI, but also encompasses other technologies, including blockchain, AR/VR, and quantum computing. Ultimately, we see a long-term benefit for organizations to respect and protect employees who may raise questions they feel are a concern to society and the long-term benefit of the company.

Assessing Disparate Impact or "Fairness"

At the beginning of 2018, our Responsible AI team sought to tackle what seemed to be one of the big problems in the ethics of AI space. Business leaders, and stakeholders across industries lacked tools to understand algorithms and thus be equipped to partake in decision making around fairness considerations in algorithms. As we approached this challenge, we looked to just two of the definitions from Arvind Narayanan's paper titled '21 fairness definitions and their politics' codified into computer science.[1]

For a technical explanation of the tool, please see "A Framework for Translating Academic Research to Application: Accenture Fairness Evaluation Tool" (appendix).

Adversarial and Trustworthy AI

Much attention has been paid to unintentional misuse of AI that can lead to problems like discrimination and bias. However, it is prudent to remember that another key component we must reckon with is the intentional attacks on an AI system, often using AI to carry out such attacks. Researchers in adversarial AI have created proof-of-concept attacks against a number of core technologies like computer vision, OCR (Optical Character Recognition) and malware detection. A brief examination of some of these highlights the methodologies and desired outcomes behind these attacks.

Computer vision

[1] The 21 definitions illustrate the challenges policymakers and standards designers may have if they attempt to codify or endorse any one system. Fairness is a challenging concept to assign a specific definition to. That is why agile management methods are critical to the future of AI.

Many recent advances in computer vision have been enabled by deep learning – from classifying image content and creating decision-making processes for self-driving cars to recognizing objects in surveillance feeds.

Image content/classification is one of the most researched areas of adversarial AI. A typical attack in this space generates an adversarial example which is given to a machine learning model. Because of manipulation, the model misinterprets the content of the image and misclassifies it.

In this way, an attacker can tailor the expected behavior of an algorithm to achieve a number of outcomes. And in self-driving car use cases, researchers have created adversarial examples that can cause accidents.

Widely used by organizations to extract text from images, OCR software is another area at risk of attack. Proof-of-concept adversarial attacks have caused OCR systems to misread the information from images that is then translated to text. Fraud use cases represent one of the broadest attack vectors (online banking apps could be exploitable).

Natural language processing (NLP)

Recent research shows that applications of deep learning in NLP are also vulnerable to adversarial attacks. Unlike images, which are usually scaled to have continuous pixel intensities, text data is largely discrete. This makes optimization for finding adversarial examples more challenging.

Adversarial examples in this space focus on inducing misclassifications through changes that maintain semantic similarity (sentences with similar meanings are close to each other) or making changes that maintain syntactic similarity (sentences are structured the same).

The objectives of these adversarial attacks are varied and could include subversive manipulation of the algorithms that determine sentiment, gather intelligence, or filter for spam and phishing.

Industrial control systems

To reduce computational complexity, many control systems make estimations and approximations. This simplification means that some interactions will not be captured in the control equations. By creating Generative Adversarial Networks (GANs) that make minor manipulations (that may go unseen by human operators) to control systems' inputs, attackers can cause unexpected behaviors that create a wide array of outcomes – from simple system degradation, to increased wear-and-tear, to catastrophic failure.

Securing AI

The majority of organizations' current investment into security is dedicated to securing the hardware and software attack surface. These include patching vulnerabilities, static and dynamic analysis of production codes, and OS hardening.

This overlooks a key point: adversarial AI targets areas of the attack surface that have never previously had to be secured, the AI models themselves. From now on, organizations need to include these in their security budgets – or risk them being exploited by attackers.

Securing an AI model requires different skills and toolsets than securing code. In large part, that's because it's impossible to test every combination of inputs for an AI model (the number of values taken by a single variable can be infinite).

Until recently, data scientists addressed this problem by using sensitivity and robustness testing. But these tests are used to test for stability on random inputs, not specific combinations of inputs engineered to trigger unexpected behavior. And they're more likely to fail to predict behavior in more complex models.

To ensure their AI models are robust enough to withstand exploitation, organizations must take advantage of adversarial AI counter-measures and emerging practices. The AI attack surface is an entirely new area of infrastructure that has to be secured. Security practices must adapt to accommodate it, including updating threat modeling processes to account for adversarial AI threats.

So how can the AI attack surface be comprehensively protected? It's a complex challenge and organizations will need to combine multiple approaches to ensure robust, secure AI such as:

- ***Rate limitation*** – by rate-limiting how quickly individuals/systems can submit a set of inputs to a system, organizations can increase the effort it takes to train their models. That's a major deterrent to adversarial attackers.
- ***Input validation*** – data sanitization focusing on what's being put into AI model while in inference mode, or while undergoing training. By making small modifications to an adversarial example, it's often possible to "break" its ability to fool a model. When performed during the training phase of the model through special techniques, this process can be used to clean poisoned training data and prevent AI trojans/backdoors.
- ***Robust model structuring –*** the structuring of machine learning models can provide some natural resistance to adversarial examples. It is worth noting that this will involve tradeoffs between the model's accuracy, its robustness, and its explainability.
- ***Adversarial training –*** if enough adversarial examples are inserted into data during the training phase, a machine learning algorithm will eventually learn how to interpret them. Adversarial examples can be generated by training a special Generative Adversarial Network (GAN).

Even though AI attack surfaces are only just emerging, organizations' future security strategies should take account of adversarial AI, with the emphasis on engineering resilient modelling structures and strengthening critical models against attempts to introduce adversarial examples.

Immediate priorities for organizations to consider include:

1. Conduct an inventory to determine which business processes leverage AI, and where systems operate as black boxes
2. Gather information on the exposure and criticality of each AI model discovered in Step 1 by asking:
    - Does it support business-critical operations?
    - How opaque/complex is the decision-making for this process?
    - Is the process exposed to the outside world?
    - Can customers create their own inputs and get results from the model?
    - Are there similar open-source models to this process?
    - What potential outcomes could an attacker drive from this model?
3. Prioritize plans for highly critical and highly exposed models:
    - Using the information gathered in Step 2, prioritize each model and create a plan for strengthening models that support critical processes and are at high risk of attack

- To support prioritization, create trade-off matrices that weigh criticality vs the risk and exposure of each model.

NIST Question #8: Technical standards and guidance that are needed to establish and advance trustworthy aspects of AI technologies.

Just like other critical technology matters such as cybersecurity and privacy, AI will require global governments to come together to establish best practices and frameworks to guide organizations. They will be key to both safety and innovation. For that, we were pleased to see countries come together for the recent **OECD Council Recommendation on Artificial Intelligence**. We encourage the United States government to adopt a similar whole-of-government approach for the other standards organizations currently working on standards such as the IEEE and ISO.

However, while cross-sectoral best practices will be useful at a high-level, industry specific and risk-based frameworks will be necessary in some high-risk cases. Through this RFI process, NIST is well-positioned to play the role of convener for stakeholders as they work to create guidance, helpful across sectors, on how to approach technical standards. It is important that NIST/the federal government consider whether NIST should also serve as a convener for non-technical frameworks.

Once the discussion reaches industry specific frameworks, we must take care to specify how the guidance applies to what specific kind of AI. And, we must ensure that best practices and guidance continue to be updated- just as the NIST Cybersecurity Framework continues to be updated taking into account new technology advances, threats, and opportunities.

We believe that effective government approaches to AI clears barriers to innovation; provide predictable and sustainable environment for business; protects public safety; and builds public trust in the technology. An example of such an approach is the UK Financial Conduct Authority's regulatory sandbox model, which has proved so successful that it has now been adopted in other countries, including Australia. As suggested in the OECD AI recommendations, we would support NIST developing a sandboxing scheme for AI technical tools, that will enable innovative businesses to test and pilot AI algorithms and tools responsibly, in a safe environment, and within a safe framework.

However, standards or "assessments" must not be used to create a checklist for ethics or fairness. Governance of algorithms must include both quantitative and qualitative measures (technical and non-technical). Quantitative measures are the empirical evidence necessary to prove AI systems are effective, fair, and transparent. Qualitative measures enable the critical thinking necessary to interpret evidence effectively. Both context and evidence are necessary to understand risks because there is no one definition of fairness, nor one understanding of sufficient transparency. Evaluations of fairness and transparency in AI systems, such as Algorithmic Impact Assessments, are a productive tool to proactively identify, mitigate and monitor these risks but they should be used to foster conversations between policymakers, regulators, and stakeholders; not to "certify" if a technology is "fair."

APPENDIX

I. A Framework for Translating Academic Research to Application: Accenture Fairness Evaluation Tool

A Framework for Translating Academic Research to Application: Accenture Fairness Evaluation Tool

Accenture Applied Intelligence

Rumman Chowdhury, Caryn Tan, Deborah Santiago, Benjamin Jones August

2018

Abstract

Fairness is a critical and much-discussed component of good algorithmic deployment. However, the definitions of fairness are numerous and diverse, and promising academic literature can be difficult to develop in practice. In this paper, we discuss the components of Accenture's Fairness Evaluation Tool and present a framework for translating academic research into practical application. We utilize the key steps of scaleability, generalizability and integrability to examine each potential solution for viability. We also discuss the concept of human centricity, and provide some concrete examples of design decisions made in the algorithmic tool development process to encourage agency and accountability via human-centric design.

1 Introduction

Algorithms are everywhere. They have become an integral part of our lives and dictate everything from which ads we see to whether or not we are approved on our mortgage, what music we are listening to and how expensive our car insurance premiums are. But what are the consequences on businesses and society as a whole if the algorithmic output is fundamentally unfair? What are the implications on society if, as is sometimes the case, men are more likely to be shown higher paid, more senior job advertisements than women?[1] Or, if re- sumes of women submitted for programming jobs were discounted because the definition of a successful employee was based on the composition of current executive pool[2]? The prevalence of algorithms in our daily lives has led to a growing concern among academics and industry alike that more needs to be done to understand and prevent the effects of systematic algorithmic bias.

Algorithms provide the veneer of technological objectivity. With data, math and programming as intermediaries, we remove ourselves from the decision-making process and presumably, remove ourselves from human-bias. However, in the context of machine learning, the machine can only learn from the training data it is provided and the constraints of the algorithm. Therefore, algorithms are not automatically fair by design. There are many ways to address algorithmic fairness. One method involves doing more to anticipate for and mitigate against "unintended consequences" of the uses of technology.

Accenture's Fairness Evaluation Tool was designed with an appreciation of the depth of contextual understanding that may be involved in an assessment of fairness. Ultimately, the weight of responsibility of this decision should not solely fall on one person; it should be a collaborative effort informed by many parties. We discuss specific human-centric design decisions that empower the user to make trade-offs and decisions, rather than outsource the outcome, and the responsibility of the consequences, to the model. The Fairness Evaluation Tool was designed for human decision-making and for simple interactive visualizations intended to explain quantitative output to a non-technical audience.

The statistical underpinning of our work is not new or novel. However, what our efforts introduce is a framework for translating academic work into integrated applications based on the concepts of scaleability, generalizability, and integrability. Thus, the subsequent sections are organized as follows: (a) a discussion of the legal and quantitative literature on fairness, (b) a proposed framework for assessing academic work for readiness in integrated applications, (c) an overview of how we applied that framework to the fairness literature to create the tool and (d) the human-centric design decisions that serve to nudge the user towards agency and encourage collaborative decision-making.

2 What is fairness?

A popular approach that has translated from the legal sphere to the quantitative academic field is the concept of using disparate impact as a red flag for determin- ing "unintended consequences" and possible unfair outcomes. This practice of anticipating unintended consequences can find historical precedent in the framework of analysis surrounding the concept of "disparate impact" found in United States jurisprudence. Under Griggs v. Duke Power Co.[3], the US Supreme Court evaluated whether an actual intent to discriminate (i.e., overt discrimination) was needed in order to conclude that an employment practice was discriminatory. The US Supreme Court found that there might be practices that are "fair in form but discriminatory in operation" even without an actual intent to discriminate. As this concept has evolved over the last 47 years, the framework is as follows in its simplest form: (1) does the employment practice have a disparate impact on protected classes of people (e.g., race, gender), (2) is there a business necessity for that practice and (3) are there any reasonable alternatives available to the employer rather than the questionable practice at hand?

Translating this legal concept to algorithmic outcomes, we find that most of the quantitative literature on fairness focuses on (1), but not (2), or (3). While the last two considerations are outside the scope of this paper, they are often raised in the context of model assessment. (2) roughly translates to "just because we can make something, should we?" and (3) asks, does the model perform better than a human? For example, in a binary classifier, we ask if the model performs better than a coin flip.

There are many ways of enforcing fairness constraints in data analytics, but not all of them can be achieved at once. Additionally, fairness often involves a compromise on model accuracy. Thus, it is important to understand what the implications of each constraint are. In this section we go into greater detail about con- cepts of fairness that are most relevant to the realm of the data scientist's work: issues around data process- ing, modelling, and model evaluation. These concepts are illustrated in subsequent sections using the publicly-available German Credit Score Dataset[4]. For simplicity, we suppose that gender is the only protected attribute when providing examples.

2.1 Quantified fairness as a measurement of disparate impact

For the purposes of the Fairness Evaluation Tool, we use the term disparate impact to describe the possible occurrence of unintended discrimination through the usage of interconnected variables. Even if gender is excluded from a model, decision-making can lead to discrimination if there are variables associated with gender in the model. Examples are variables such as salary and profession, which have different distributions for each gender. It is also helpful to consider the less serious example of shoe size, because it is a good example of non-obvious associations. After a model is built with a protected variable excluded from the build, we know some form of discrimination may have occurred if the probability of an outcome is not the same for different values of the protected variable.

2.2 Quantified fairness as Predictive Parity, Equal Opportunity, and Equal Leniency

If a model is well-calibrated, then the classifier exhibits predictive parity if it obtains similar predictive values for different groups within a protected variable (for example, the predicted value is similar for males and females). *Equal opportunity* means that the true positive rate (TPR) is equal across the protected groups, where the TPR is defined by Verma and Rubin as the fraction of positive cases correctly predicted to be in the positive class out of all actual positive cases[5]. It is often referred to as sensitivity or recall as it represents the probability of the truly positive subject to be identified as such. In our example, a true positive is a person who paid back their loan and for whom it was predicted that they would do so. In this context, a difference in TPR is unfair because it means that the rate at which the model predicts the individuals who were loan-worthy is different between subgroups. *Equal leniency* is also referred to as predictive equality. Leniency is a measurement of False Positive Rates (FPR), or, the fraction of negative cases incorrectly predicted to be in the positive class out of all actual negative cases.

From a measurable and applied perspective, a data scientist needs to understand the context of the model and which of these fairness quantifiers is the most rational. Leniency, or false positive rates, can often be a better measure of fairness than others. True negative rates are not possible to measure in many cases, as we cannot determine the counterfactual outcome. While equal opportunity is enforceable, it is difficult to optimize. To ensure equality in the wild for our example, we may have to enforce a very low credit approval rate, which is not an optimal business outcome. From a business perspective, one may choose to optimize for leniency. A leniency approach is more strict on subgroups that are

favoured, and increases the level of acceptance for the groups that are traditionally less favoured. The bank can reduce cost with fewer loan defaults from the favoured group, thereby reducing overall cost, and increase revenue by allowing more people to receive (and pay back) loans who may not have otherwise received them.

2.3 Fairness as collaboration

In defining fairness, we caution the development of a tool that relies on pure algorithmic solutioning. This is because pure algorithmic solutioning often lacks the requisite context that enables informed decision-making. We also appreciate that the responsibility of identifying an "ideal" level of fairness may be a heavy burden to bear for an individual data scientist, and therefore we approach fairness as a collaborative effort. In particular, an organization's definition of fairness should be a function of their core business values, industry-specific requirements, and the definition of success for the product.

Human-centric design is a popular, and possibly overused, term in the technology product space; here, human-centricity allows for the collaboration that is required for true contextual awareness. For the Fairness Evaluation Tool, we define human centricity as enabling human agency, accountability and understandability. In subsequent sections, we will describe the human-centric nudges included to inspire ownership over the outcome and collaborative decisioning.

3 Translating academic work to application

Our tool prototyping began with a broad range of academic literature on fairness. However, academia is often focused on exploring the boundaries of what can be achieved, and it can often be difficult to translate academic research into a product offering that can be applied to industry. As a result, we focused on the following criteria for assessment: (1) scaleability, (2) generalizability, (3) integrability.

For technology product designers, scaleability is the critical component in the creation of a new offering. Scaleability is defined as: the ability of a process, network, software or organization to grow and manage increased demand. In other words, it is critical that we are able to provide consistency and accuracy in results, and have an expectation that this process will not impede workflow.

Generalizability asks whether the statistical underpinnings of the code is something a general data scientist can understand and execute with a high level of understanding with a minimal amount of training. Highly specialized academic literature may require a skill set that is beyond the capability of data science and AI teams at many companies, particularly those outside of the traditional technology fields. Even in more algorithmically sophisticated industries (e.g. banking), we may find that the fairness literature requires niche capability they do not possess.

Finally, integrability appreciates that ethical data and algorithmic practices are often seen as an impediment to progress. A commonly heard phrase is "regulation stifles innovation"– and ethical assessments are consid- ered part of this regulation. To address this, the Fairness Evaluation Tool mandate was "flexibility by design". Each of the models we used is assessed for ease of integration, and our development team is focused on creating seamless pipelines from data to deployment. As a result, the model output needs to exist in a format that is easily integrated to a variety of cloud-based or on-premise machine learning solutions.

As an example of an application of this framework that led to the elimination of a fairness procedure, we examined the counterfactual fairness paper[6] for the Fairness Evaluation Tool. Kusner et al define counterfactual fairness in the context of decision making as "a decision that is fair towards an individual if it is the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group". The authors point out that enforcing equal opportunity at the modelling stage will not meaningfully protect equal opportunity if the individuals in the training data have not experienced equal opportunity. However, this academic research, while groundbreaking, did not meet the application framework criteria. Counterfactual fairness requires an understanding of causal relationships and the ability to map these linkages in a meaningful manner to understand latent causes of bias. As model complexity increases, the difficulty of mapping these linkages increases exponentially. This skill is not taught as part of a standard data science practice, and the learning curve may be steep. Finally, the process may take longer than is acceptable for a standard development process. A cost-benefit analysis here is key to the success of applicable research.

Using this criteria, our examination of the literature results in a tool with three parts: i) a data investigation tool that examines the hidden impact of sensitive variables (Section 6.1), ii) a tool to look at disparate im- pact of the model outcome (Section 6.2) and iii) a tool to enforce equalized leniency for classification models (Section 6.3).

4 The analytical journey

Often overlooked, the analytical journey undertaken by data scientists is a key first indicator of bias in a given model. Figure 1 compartmentalises the data journey, highlighting some fairness considerations that can be accounted for at each stage. Olteanu et al.[7] provide a taxonomy of biases that can occur. Here, we step through the analytics workflow, drawing attention to a few examples of biases that can arise. We recommend a close reading of Olteanu et al. to understand the complete picture.

At the point of ideation: Is the group of people tasked with brainstorming relatively homogeneous? If so, they may not have an adequately diverse set of viewpoints between them. This is particularly troubling if the analysis concerns a population of people that include individuals very different to the brainstorming decision-makers.

Data collection: Is the collected sample representative of the population it is supposed to represent? If you're trying to build a model to predict whether a person will default on their loan, and the sample is made up of people who were granted a loan based on criteria such as credit score, then your model will not be predictive in the population at large, but merely in the subpopulation of people who had a credit score exceeding a certain threshold.

Data processing: There are many issues that can arise here. One example concerns missing values. If a choice is made to work with "complete cases" only, then any observations with missing data will be excluded from the analysis. Yet this missingness could be systematic, so that a subpopulation is systematically underrepresented after the data cleaning. Data aggregation can also cause bias because reducing the granularity of the data may obscure crucial differences between subpopulations.

Analysis and evaluation: Depending on the model that is chosen to represent the data, very different results can be achieved. It is important to compare several models in terms of predictive performance in subgroups defined by protected characteristics such as gender. In Section 6 we highlight several methods evaluating and adjusting for differences in model accuracy between these subgroups.

5 Human centricity and SGI framework applied to the Fairness Evaluation Tool

First, in the data investigation tool, we allow the user to select their sensitive variables. In the disparate impact tool, we allow the user to select their desired level of repair, and visualize both the level of repair as well as the accuracy shift. In the predictive parity tool, we allow the user to select the leniency (false positive rate) and illustrate the cost associated with this value. In each case, rather than pre-select or optimize an output, we assume the user has the contextual knowledge to select the right value for their trade-off function.

Our visuals are designed to invite conversation. For example, we choose to illustrate the cost – not accuracy level - associated with the predictive parity value. Provided a 2x2matrix of cost for false positive, true positive, false negative, and true negative outcomes, we calculate a monetary value of the adjustment. In doing so, we enable the data scientist to have an outcomes-related discussion with key stakeholders. In the example of the data used for testing the tool, we model the monetary cost of predictive parity, and illustrate that some levels of adjustment lead to lower costs for the bank, due to a decline in defaults and additional opportunities for those previously denied.

6 The Fairness Evaluation Tool

For each concept of fairness outlined in Section 2 (minus counterfactual fairness), we have implemented methods for quantifying and adjusting both the data and the modelling process. This includes evaluating the fairness-accuracy trade-off inherent in each process. We propose a platform for integrating the considerations of model accuracy and its cost implications with those of fairness. We also illustrate how each fairness component fits in with the typical data science workflow.

6.1 Mutual information

To get a better understanding of how the variables interrelate, bi-variate analyses of each combination of variables was carried out. Mutual information was chosen as the measure of inter-variable dependence. Brown et al.[8] provide a short introduction to this metric. The mutual information between two variables, X_1 and X_2, tells us how much knowing the value of one variable, say X_1, informs us about the value the other, X_2, might

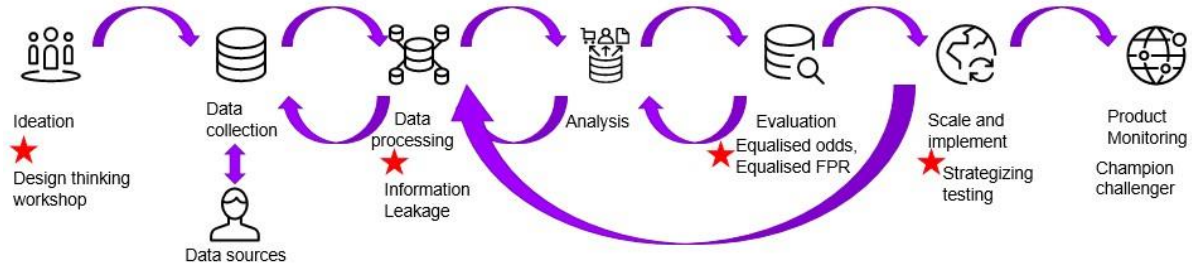


Figure 1: The stages of the analytical process.

take. It is calculated as

$$\mathrm{MI}(X_1;\, X_2) = \sum_{x_1, x_2} p(x_1, x_2) \log \frac{p(x_1, x_2)}{p(x_1)p(x_2)}, \quad (1)$$

where x_1 and x_2 each represent the values taken by the X variables. Note that if the variables are independent, $p(x_1, x_2) = p(x_1)p(x_2)$ so that the MI value will be zero. In reality, it will be close to, but not exactly, zero if the variables are independent.

The maximum possible value of MI is achieved if the two variables are completely dependent, so that knowing one tells you what value the other will take. However, MI is not just a function of dependence, but also depends of the number of different values taken by each variable and the marginal distribution of each, so that the maximum possible value varies between variable combinations.

MI can only be applied to discrete variables. We discretized the continuous variables by binning each into a maximum of five bins with an equal number of observations, N , in each bin. Figure 2 displays the normalized MI values for each variable combinations. To test the significance of these values, calculate the G-statistic as $2 * N * MI(X_1; X_2)$. It follows the χ^2 distribution[9].

6.2 Disparate impact

6.2.1 Evaluating disparate impact

Feldman et al.[10] propose a model based approach to identifying disparate impact by using the concept of balanced error rate (BER). If $\boldsymbol{X} = (X_1, \ldots, X_n)$ are the non-sensitive predictors of the German Credit Data and G is gender, build a model, $f(X)$ to predict G from X. Then the BER is defined as:

$$\mathrm{BER}(f(X), G) = \frac{P(f(X) = m \mid G = m)}{}$$

Then G is said to be E-predictable from X if $\mathrm{BER}(f(X), G) \leq E$. And it is "$E$-fair" if the BER exceeds this threshold. See Feldman for a full description of the method. As this method is model dependent, it

must be kept in mind that if the "right" classifier is not applied, the BER will not be accurate. In this study, a hinge-loss Support Vector Machine (SVM) algorithm was applied, in line with the approach taken by Feldman et al.

Where bias is not corrected for, or not completely corrected for, it is possible for this bias to become amplified in the model build process. Bias amplification identification was proposed by Zhao[11] to evaluate the change in disparity between two groups in terms of outcome. We did not focus on this method in the Data Study Group, but wish to highlight its existence. In summary, first calculate the maximum likelihood probability of default based on the observed outcome for each subgroup, $P(D = 1 \mid G = m)$ and $P(D = 1 \mid G = f)$. Then compare this to the classifications output from the model, $P(\hat{D} = 1 \mid G = m)$ and $P(\hat{D} = 1 \mid G = f)$. If

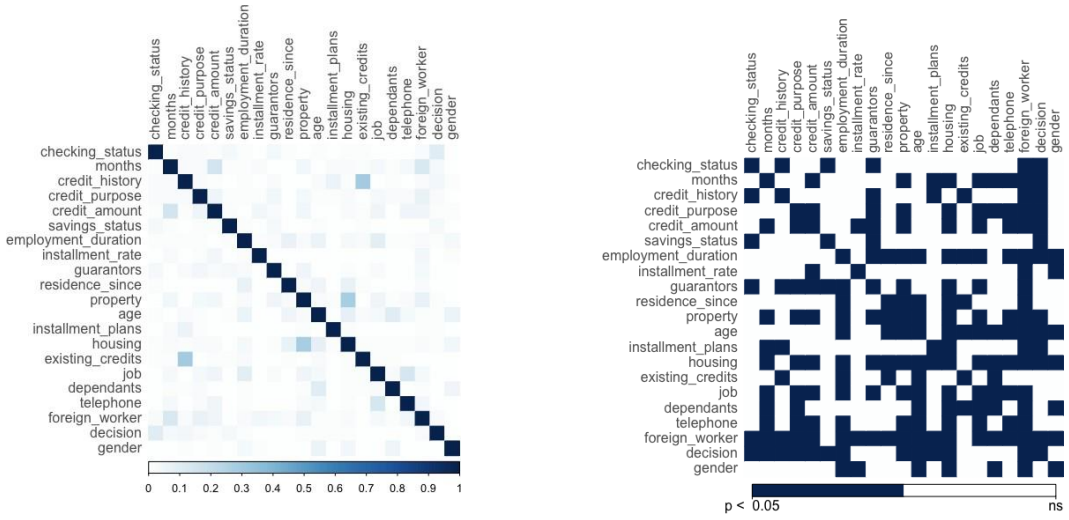
$$\frac{+P(\hat{f}(X) = f \mid G = F)}{2} \quad (2)$$

there is a difference between the values, this suggests that bias amplification has occurred.

6.2.2 Correcting for disparate impact

Feldman et al. state that their method removes all information leakage that leads to disparate impact while preserving the rank[10]. In our case, this is the rank of individuals in terms of credit-worthiness (we did not verify this in our experiments). They propose several approaches for partially "repairing" the data so that the disparate impact is reduced. The complete removal of disparate impact is cautioned against because it can lead to a significant reduction in model accuracy.

Our experiments based on the credit data show the effect of repairing the data with respect to protected variable gender, Figure 3. This is a high-dimensional data set, so, to enable a meaningful visualisation of how the classification quality changes with adjustment for



(a) NMI values

(b) Significance testing

Figure 2: Normalized mutual information for each combination of variables in the German Credit Data in plot (a). In plot (b), dark blue squares indicate a statistically significant relationship between two variables at the 0.05 level.

disparate impact, we projected the dataset into a two-dimensional embedding. The axes are chosen as the normal vectors of the separating planes of classifiers trained on reconstructing the protected attributes and predicting the credit rating, respectively. So the x-axis equals the decision value of a linear classifier for reconstructing the protected attribute, and the y-axis equals the decision value of a linear classifier for predicting the credit rating.

It is advisable to explore the fairness-accuracy trade-off of this method for various degrees of repair, ranging from none to complete, before finalising the extent of repair. Any compromise made on the accuracy of the model will impact a bank's risk profile. So, similarly, the fairness-cost trade-off should also be made clear. See Figure 4 for what this visualisation might look like.

If achieving the desired standard of fairness requires too great a compromise in terms of model accuracy, the data collection process should be scrutinised. A completely new data set could be the more appropriate solution.

6.3 Equalized leniency

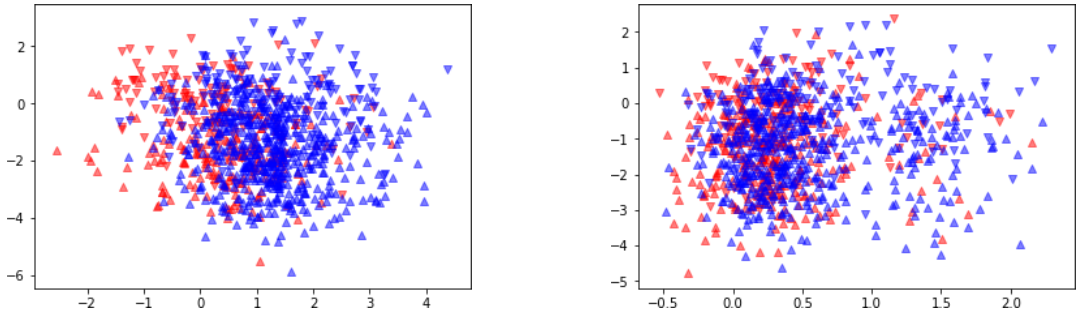
Several methods have been proposed[12,13] to impose equal treatment amongst subgroups, say male/female gender, of a data set by adjusting the classification threshold on model output. These methods can easily be inserted into a workflow in so far as they are agnostic to the type of model that has been used to generate the output. The only requirement is that the model outputs a continuous prediction. In this study we focus on probability outputs, but it is straightforward to extend this approach to other continuous model outputs.

A disadvantage of this approach is that it can be naive. Consider the problem of information leakage described in Section 6.1. Correcting for differences with respect to a protected variable will not address the issue completely if predictor variables are related to it.

A logistic regression model with elastic net is applied to the credit data to obtain predicted probabilities of being given a loan. We then visualise any divergence in model performance between the categories of the protected variable using ROC curves. ROC curves permit a comparison of the TPR and FPR for every possible threshold on the model output. The ROC curve in the Figure 5 (a) shows that there is a difference between the two age groups, under 30 years and 30 or older, in terms of model accuracy.

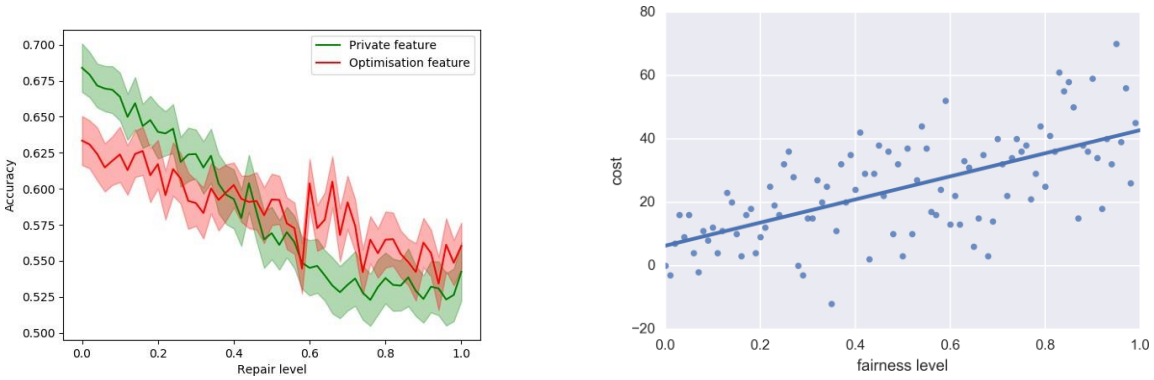
It is possible to re-threshold the probabilities output by the predictive model in order to achieve equal FPR and/ or TPR in both subgroups. However, Chouldechova[14] has shown that the TPR, FPR, and PPV cannot be equalised at the same time using classification calibration techniques:

$$\mathrm{FPR} = \left(\frac{p}{1-p}\right)\left(\frac{1-\mathrm{PPV}}{\mathrm{PPV}}\right)(1-\mathrm{FNR}) \qquad (3)$$



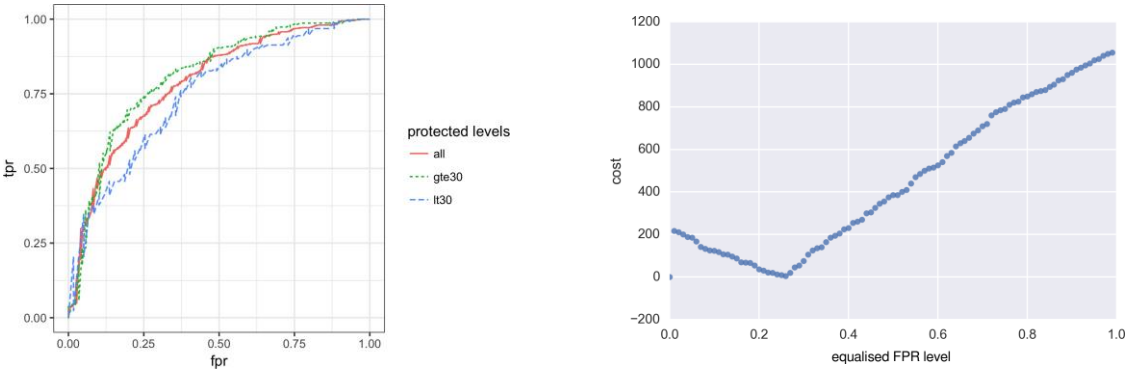
(a) No corrections (b) Corrected

Figure 3: The plots show the data set (a) before it is corrected for disparate impact, and (b) after the correction has been applied. It can be seen that data points of the same category clustered towards the edges are better interspersed with correction applied, and thus anonymized



(a) fairness-accuracy trade-off (b) fairness-cost trade-off

Figure 4: Visualising the trade-offs that are involved in enforcing fairness constraints in disparate impact. Cost analysis is based on a cost of five units for a true positive and one unit for a false positive.



(a) ROC curve (b) Cost analysis

Figure 5: (a) compares the ROC curves of subgroups before classification calibration is applied. (b) shows the fairness-cost trade-off for different levels of equalised FPR.

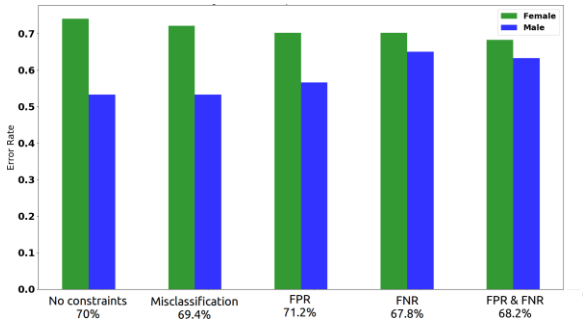


Figure 6: A comparison of overall accuracies (listed beneath each pair of bars) for the unconstrained approach and model calibrations which enforce parity amongst subgroups of misclassification rate, FPR, FNR, and FPR & FNR.

Chouldechova notes that if the PPV is kept the same across subgroups but the prevalence, *p*, differs between groups, equal FPR and TPR across subgroups cannot be achieved.

A cost analysis of the impact of different equalised FPR thresholds is provided in Figure 5 (b). Depending on what the primary fairness concerns are, an alternative plot can be generated to compare, say, the TPR-PPV trade-off instead of the TPR-FPR of Figure 5.

7 Conclusion

Incorporating a definition of fairness from academia into the data science workstream of corporate applications is challenging. In this paper, we have presented a methodology for translating ethical AI research into disruptive, industry-standard applications, using the Accenture Fairness Evaluation Tool as a use-case, with human-centricity at the core. We break this down into three key areas: scaleability, generalizability, and integrability, and discuss how each is relevant to the responsible AI field.

During this process, we discovered some key learnings. When correcting for equalized leniency in Section 6.3, we found the interesting relationship seen in Figure 6. As we enforce parity amongst subgroups of misclassification rate, we obtain a higher overall accuracy. This is a welcome but unexpected result. An explanation of this may come by considering, in the context of the German Credit Data, that certain subgroups are being given an opportunity to access credit that they were not other- wise able to. As this technique is applied in future ap- plications, we look to investigate this finding to create a reliable benchmark for the fairness-accuracy tradeoff.

Future implementations of this tool will take into account the limitations of infra-marginality with regard to the application of predictive parity to marginalized groups. This issue can be addressed with a guardrails approach to compare differing risk distributions.[15]

It should be made clear that the Fairness Evaluation Tool is limited in the number of variables that can be corrected simultaneously. Correcting for one variable can be achieved as discussed, but initial findings suggest that correcting for disparate impact on multiple variables leads to an inadequate compromise on predictive quality. Instead, a decision must be made as to which variable is most "impactful" given the context. This provides further evidence for the conclusion that fairness cannot be decided solely by a tool; rather the tool should drive a larger discussion around accountability, governance and ethics in algorithmic decision making.

Acknowledgements

Accenture Applied Intelligence would like to thank the Alan Turing Institute for their initial work on the Fairness Evaluation Tool. We would also like to thank the following: Peter Byfield, Paul-Marie Carfantan, Omar Costilla-Reyes, Delia Fuhrmann, Jonas Glesaaen, Qi He (Katherine He), Andreas Kirsch, Julie Lee, Mohammad Malekzadeh, Esben Sørig, Emily Turner, and Dang Quang Vinh for their invaluable work and collective insights.

References

[1] A. Datta, M. C. Tschantz & A. Datta. *CoRR* **abs/1408.6491.** arXiv: 1408.6491. http://arxiv.org/abs/1408.6491 (2014).

[2] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama & A. Kalai. *CoRR* **abs/1607.06520.** arXiv: 1607.06520. http://arxiv.org/abs/1607.06520 (2016).

[3] U. S. Court. *Griggs v. Duke Power Co.* 1971.

[4] D. Dua & E. Karra Taniskidou. *UCI Machine Learning Repository*

[5] S. Verma & J. Rubin. in *Proceedings of the International Workshop on Software Fairness* (ACM, Gothenburg, Sweden, 2018), 1–7. ISBN: 978-1-4503-5746-3. doi:10.1145/3194770.3194776. http://doi.acm.org/10.1145/3194770.3194776.

[6] M. J. Kusner, J. R. Loftus, C. Russell & R. Silva. in *NIPS* (2017). arXiv: 1703.06856. http://arxiv.org/abs/1703.06856.

[7] A. Olteanu, C. Castillo, F. Diaz & E. Kiciman. *SSRN Electronic Journal,* 1–44. ISSN: 1556-5068 (2016).

[8] G. Brown, A. Pocock, M.-J. Zhao & M. Lujan. *Journal of Machine Learning Research* **13,** 27– 66. ISSN: 01678655 (2012).

[9] B. Woolf. *Annals of Human Genetics* **21,** 397– 409. ISSN: 14691809 (1957).

[10] M. Feldman, J. Moeller & C. Scheidegger. *arXiv1412.3756,* 1–28 (2015).

[11] J. Zhao, T. Wang & M. Yatskar. in *EMNLP* (2017).

[12] M. Hardt, E. Price & N. Srebro. *Computing Research Repository.* https://arxiv.org/pdf/1610.02413.pdf (2016).

[13] M. B. Zafar, I. Valera, M. G. Rodriguez & K. P. Gummadi. *AISTATS* **54** (2017).

[14] A. Chouldechova, 1–17. ISSN: 2167-6461 (2017).

[15] S. G. Sam Corbett-Davies. *Computers and Society.* arXiv: 1808.00023. https://arxiv.org/abs/1808.00023 (2018).