Assessing AI Fairness in Finance

Lachlan McCalman and Daniel Steinberg, Gradient Institute
Grace Abuhamad and Marc-Etienne Brunet, ServiceNow
Robert C. Williamson, University of Tübingen
Richard Zemel, University of Toronto

If society demands that a bank's use of artificial intelligence systems is "fair," what is the bank to actually do? This article outlines a pragmatic and defensible answer.

Artificial intelligence (AI) systems are becoming ubiquitous in a diverse and ever-growing set of decision-making applications, including in the financial sector. AI systems can make consequential decisions at a speed and volume not possible for humans, creating new opportunities to improve and personalize customer service but also increasing the scale of potential harm they can cause if they are misdesigned.

AI systems unfairly discriminating against individuals by their race, gender, or other attributes is a particularly common and disheartening example of this harm. For example, soon after launching its credit card partnership with Goldman Sachs in 2019, Apple had to investigate its system for gender bias. This bias, if left unchecked, could have limited women's access to credit, harming those potential customers and increasing risks of regulatory noncompliance for the business.[1] However, there is no simple solution to preventing these kinds of incidents: helping AI live up to its promise of better and fairer decision making is a tremendous technical and social challenge.

One of the key design mistakes behind harmful AI systems in use today is an absence of explicit and precise ethical objectives or constraints. Unlike humans, AI systems cannot apply even a basic level of moral awareness to their decision making by default. Only by encoding mathematically precise statements of our ethical standards into our designs can we expect AI systems to meet those standards.

Digital Object Identifier 10.1109/MC.2021.3123796
Date of current version: 12 January 2022

EDITOR **HSIAO-YING LIN**
IEEE Member; hsiaoying.lin@gmail.com

Technical work to develop such ethical encodings is burgeoning, with much of the focus on the fairness of AI systems in particular. This work typically involves developing mathematically precise measures of fairness suitable for designing into AI systems. Fairness measures use the system's data, predictions, and decisions to characterize its fairness according to a specific definition (for example, by comparing the error rates of the system's predictions between men and women). The exercise of defining fairness in mathematical terms has not "solved" fairness but rather surfaced the complexity of the problem at the definitional stage. There now exists a panoply of fairness measures, each corresponding to a different notion of fairness and potentially applicable in different contexts.

Parallel to the work of encoding ethical objectives mathematically is a broader social effort to develop principles and guidelines for ethical AI. These aim to help the designers, maintainers, and overseers of AI systems recognize and ameliorate ethical risks. Governments, corporations, and other organizations have released hundreds of such frameworks in the last few years, many with common themes like the importance of explanations of an AI system's decisions, the need to provide mechanisms for redress when errors occur, and the need to understand and minimize avoidable harms caused by the system. For example, the National Institute of Standards and Technology released a proposal last year to identify and manage bias using a three-stage approach.[2]

However, a gap remains between the technical efforts and the broader design principles. Designers building AI systems have access to principles, on the one hand, and mathematical tools, on the other, but little guidance about how to integrate these two resources and build a system that utilizes them in consequential settings.

THE FEAT PRINCIPLES AS A STARTING POINT

Financial services institutions (FSIs) manage billions of dollars' worth of transactions per day and are increasingly adopting AI solutions as part of this business, including for determining loan and credit card approvals, conducting marketing, and detecting fraudulent behavior. In particular, the Massachusetts Institute of Technology *Technology Review Insights* found that businesses in the Asia-Pacific region are quicker to adopt AI systems than any other part of the world.[3] The scale and importance of these systems to FSI's daily operations means that if they are misdesigned, they can create reputational, operational, and legal risks for businesses and unnecessary harms to customers.

To begin addressing the ethical risks of AI decision making in finance and in doing so encourage AI adoption, the Monetary Authority of Singapore (MAS) released principles for responsible AI in the finance industry.[4] These "FEAT Principles" (Fairness, Ethics, Accountability, Transparency) were developed in partnership with Singaporean and international financial institutions and AI experts, known as the Veritas Consortium,[5] and describe aspirational ethical properties that an AI system would have, such as not systematically disadvantaging individuals, or groups, without justification.

While appearing simple, these principles contain within them complex and value-laden questions such as when a group or individual is being "systematically disadvantaged" and what data count as "relevant" for a particular application. Like the concept of fairness itself, these questions have no single uncontested answer, nor one that is independent of ethical judgment. Nor do the principles provide guidance for which (if any) of the myriad fairness measures developed may be appropriate to use to specify unjustified systematic disadvantage or unintentional bias.

FROM PRINCIPLES TO GUIDANCE: FEAT FAIRNESS ASSESSMENT METHODOLOGY

Since releasing the FEAT Principles, MAS and the Veritas Consortium have worked with teams of experts to develop implementation guidance. In January 2021, they released two white papers that detailed the first step in that implementation: a methodology for assessing AI systems for alignment with the FEAT Fairness principles[6] (with the other principles relating to Ethics, Accountability, and Transparency being tackled in a later phase) and a set of detailed case studies illustrating the application of the methodology to credit scoring and customer marketing systems.[7] We led the development of the methodology and the case studies as part of the core authorship team. The methodology comprises a set of questions (and accompanying guidance) answered by the people responsible for the AI system. Their answers go to an independent set of assessors that judge the system's alignment with the FEAT Principles.

The design of the methodology had to accommodate two critical but conflicting requirements: It had to be generic enough to be applicable across a whole industry and applicable to systems with different purposes, but specific enough to be useful and implementable by practitioners who may not be experts in algorithmic fairness or ethics. The final design of the methodology tries to balance these competing requirements with three key design pillars: asking users to stake their ethical

claim, focusing on the harms and benefits of the system, and scaling the assessment to the system risk.

Asking System Owners to Stake a Claim

The first design pillar of the methodology is that it asks system owners to stake a claim on what they believe the fairness objectives of the system should be. Any assessment that can be applied to different AI systems cannot itself mandate specific notions or measures of fairness, such as the exact circumstances that constitute unjustified systematic disadvantage (see FEAT Principle 1). Different measures of fairness imply different ethical stances, and no methodology could hope to enumerate the right choice in every situation, nor impose a particular choice that aligns the designer's (or a particular community's) ethical stance.

In philosophical literature, fairness is known as an "essentially contested" concept. While the general notion of fairness is commonly understood, different people will have different ideas about exactly what is fair in a particular context. This also applies to the selection of precise fairness objectives that can be encoded into an AI system. For example, in a hiring scenario, both the application of gender quotas to remove the effects of past and current discrimination, as well as blind hiring in which the gender of applicants is obscured, are just two of many conflicting versions of fair hiring. Each of these approaches entails a different hiring process and will produce different results. Each has proponents and detractors, both with reasoned arguments that may depend on the details of the particular situation and the necessary choice of a baseline against which to compare. Deciding on a particular fairness measure for an AI system is akin to selecting one of these approaches to fair hiring; the use of a particular mathematical measure of fairness implies a specific set of ethical values and priorities.

Imposing particular fairness measures on a whole class of AI systems would certainly be ignoring the unique circumstances and context of each system as well as the ethical preferences of the people responsible for it. Therefore, the set of fairness measures can only be decided at a per-system level. Because no jurisdiction has yet developed regulation that mandates certain measures in certain circumstances (which may not even be possible or advisable), it must be the people responsible for that system that decide how its fairness should be measured.

The FEAT Fairness Assessment Methodology is built around this idea of the system owners "staking a claim" by stating their fairness objectives and how they're measured, preferably at design time. The assessment then asks them for evidence to convince an independent assessor that the system meets these objectives. This approach separates the question of "what is fair in this situation?" from the question of "does this system operate in accordance with its stated fairness objectives?" An expert can answer the second question with the output of the methodology. By sharing parts of the assessment with people affected by the system, independent ethics panels, external regulators, or the wider public, the answer to the first question can also be examined and critiqued.

Focusing on Harms and Benefits

The second design pillar of the methodology addresses the problem that to be useful, the methodology cannot simply offload all of the work of developing and measuring fairness objectives and constraints onto the users. To help in this task, the methodology asks system owners to analyze the harms and benefits that the system may create, and the different individuals and groups that it may impact. Once FSIs have identified these, they can develop fairness measures from them by estimating how these harms and benefits are distributed across the population. The resulting fairness measures may have already been developed in the literature or could be novel and specific to the system.

This approach inverts the common question of which fairness measure to choose? for an AI system: instead, it asks system owners to first decide who the system impacts and under what circumstances (noting that these choices also involve ethical judgment). Specific fairness measures can then be derived from the harms, benefits, and impacted people with guidance from the methodology.

However, understanding and developing measures for a system's impact is likely a substantial undertaking, especially when the impact may be indirect or difficult to measure. For consequential systems this effort is paramount, but for the potentially hundreds of small, proof-of-concept or research-style models used within an FSI, performing a full assessment may be an impossible workload.

Scaling for Risk

The third and final design pillar of the methodology addresses the workload involved in assessing the hundreds of AI systems in a large organization. It specifies that systems with greater risk, for example, that affect many people or that make consequential decisions, should be assessed in greater detail.

FSIs typically already undertake these kinds of risk-scaled model assessments but with a focus on financial harms. The methodology is designed to be incorporated into these processes, adding considerations of fairness risks for customers. The way it is integrated is not prescribed owing to how differently FSIs organize their internal processes, however, the methodology does make suggestions based on common model risk management approaches within FSIs.

NEXT STEPS TOWARD IMPLEMENTATION

To ensure that the final version of the assessment methodology was indeed useful and practical to implement, we applied it to a number of real and synthetic AI systems, releasing these as accompanying case studies.[7] The case studies focused on two application areas in which AI systems are commonly deployed: customer marketing and credit scoring.

Both use cases have fairness risks traditionally and deploying AI systems in these cases can amplify these risks or introduce new ones. Credit scoring has faced risks such as the consequential impact of decisions and managing evidence of historical discrimination. Marketing also risks harming vulnerable people when targeting products, such as promoting high-interest credit cards to compulsive spenders. For both credit scoring and marketing systems, the scalability and consistency of AI decision making exacerbate potential for systematic harm to groups of customers over others.

The Veritas Consortium has now reviewed assessment methodology, and members are likely to implement the assessment methodology internally. In 2021, work continued on assessments and guidance for the other FEAT Principles (the "Ethics, Accountability, and Transparency" parts) and case studies for AI systems used in insurance. These concepts are not independent of fairness, so we will likely see iteration of the fairness methodology and integration into a single, holistic assessment.

We hope that, while being voluntary, FEAT Fairness assessments will become common practice in the finance industry and that regulators around the world will study them carefully to stimulate and inform future guidelines and regulation. We also hope that institutions begin to publish some or all of their FEAT Fairness assessments, giving the wider community an ability to understand, and voice opinions on, these systems that make consequential yet currently opaque impacts on many people's lives.

ACKNOWLEDGMENTS

The FEAT Fairness Assessment Methodology was prepared and issued by the MAS, HSBC, UOB, EY, Element AI, Gradient Institute, and IAG Firemark Labs. Authors McCalman, Abuhamad, Steinberg, and Brunet served as part of the core team. Authors Williamson and Zemel served as academic advisors for the development of the methodology. This article reflects the authors' views, and the Monetary Authority of Singapore is not responsible for the information contained.

REFERENCES

1. T. Telford, "Apple Card algorithm sparks gender bias allegations against Goldman Sachs," *The Washington Post*, 2019. [Online]. Available: https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/
2. R. Schwartz, L. Down, A. Jonas, and E. Tabassi, "A proposal for identifying and managing bias in artificial intelligence draft," NIST Special Publication 1270, 2021. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270-draft.pdf
3. "The global AI agenda: Asia-Pacific," MIT Technology Review Insights, 2020. [Online]. Available: https://www.technologyreview.com/2020/04/23/1000335/the-global-ai-agenda-asia-pacific/
4. "Principles to promote fairness, ethics, accountability and transparency (FEAT) in the use of artificial intelligence and data analytics in Singapore's financial sector," Monetary Authority Of Singapore, 2018. [Online]. Available: https://www.mas.gov.sg/publications/monographs-or-information-paper/2018/FEAT
5. "MAS partners financial industry to create framework for responsible use of AI," 2019. https://www.mas.gov.sg/news/media-releases/2019/mas-partners-financial-industry-to-create-framework-for-responsible-use-of-ai
6. "Veritas document 1 FEAT fairness principles assessment methodology," Veritas Consortium, 2020. [Online]. Available: https://www.mas.gov.sg/-/media/MAS/News/Media-Releases/2021/Veritas-Document-1-FEAT-Fairness-Principles-Assessment-Methodology.pdf
7. "Veritas document 2 FEAT fairness principles assessment case studies," Veritas Consortium, 2020. [Online]. Available: https://www.mas.gov.sg/-/media/MAS/News/Media-Releases/2021/Veritas-Document-2-FEAT-Fairness-Principles-Assessment-Case-Studies.pdf

LACHLAN MCCALMAN is chief practitioner at the Gradient Institute, Canberra, 2006, Australia. Contact him at lachlan@gradientinstitute.org.

DANIEL STEINBERG is a principal researcher at the Gradient Institute, Canberra, 2006, Australia. Contact him at dan@gradientinstitute.org.

GRACE ABUHAMAD is an applied research scientist at ServiceNow, Montreal, H2S 3G9, Canada. Contact her at grace.abuhamad@servicenow.com.

MARC-ETIENNE BRUNET is an applied research scientist at ServiceNow, Montreal, M5G 1M1, Canada. Contact him at marc-etienne.brunet@servicenow.com.

ROBERT C. WILLIAMSON is the W3 Professor in Foundations of Machine Learning Systems in the Faculty of Mathematics and Natural Sciences, University of Tübingen, Tübingen, 72074, Germany. Contact him at bob.williamson@uni-tuebingen.de.

RICHARD ZEMEL is a professor of computer science at the University of Toronto and the cofounder and director of research at the Vector Institute, Toronto, M5G 1M1, Canada. Contact him at zemel@cs.toronto.edu.