Beyond Model Interpretability: Socio-Structural Explanations in Machine Learning

ANDREW SMART, Google Research, USA
ATOOSA KASIRZADEH, Google Research, USA

What is it to interpret the outputs of an opaque machine learning model? One approach is to develop interpretable machine learning techniques. These techniques aim to show how machine learning models function by providing either model-centric local or global explanations, which can be based on mechanistic interpretations (revealing the inner working mechanisms of models) or non-mechanistic approximations (showing input feature-output data relationships). In this paper, we draw on social philosophy to argue that interpreting machine learning outputs in certain normatively-salient domains could require appealing to a third type of explanation that we call "socio-structural" explanation. The relevance of this explanation type is motivated by the fact that machine learning models are not isolated entities but are embedded within and shaped by social structures. Socio-structural explanations aim to illustrate how social structures contribute to and partially explain the outputs of machine learning models. We demonstrate the importance of socio-structural explanations by examining a racially biased healthcare allocation algorithm. Our proposal highlights the need for transparency beyond model interpretability: understanding the outputs of machine learning systems could require a broader analysis that extends beyond the understanding of the machine learning model itself.

ACM Reference Format:
Andrew Smart and Atoosa Kasirzadeh. 2024. Beyond Model Interpretability: Socio-Structural Explanations in Machine Learning. 1, 1 (September 2024), 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

> In order to formulate a learning theory of machine learning, it may be necessary to move from seeing an inert model as the machine learner to seeing the human developer—along with, and not separate from, his or her model and surrounding social relations—as the machine learner.
> - Reigeluth & Castelle [55]

The past decade has seen massive research on interpretable machine learning (ML).[1] Here is a rough restatement of the goal of interpretable ML research program: many ML models are *opaque* in that even the expert humans cannot robustly understand, in non-mathematical terms, the reasons for why particular outputs are generated by these models [31, 42, 66]. To overcome this opacity, various model-centric techniques have been developed to interpret their outputs. These techniques are diverse. They range from producing counterfactual explanations or heatmaps that offer insights into how changing inputs affect outputs [28, 41, 46], to interpreting the inner workings of the model by probing patterns of neuron activations or attention mechanisms [10, 15, 48].[2]

Despite these advancements, ML interpretability remains a contentious and ambiguous topic in the scientific community, lacking a universally accepted scope and definition [11, 13, 38, 45]. This ambiguity complicates the evaluation and regulation of opaque ML systems, raising questions about what constitutes sufficient interpretation and how it

[1] For the purposes of this paper, we use "interpretable ML," "explainable ML," "interpretable AI," and "explainable AI" interchangeably.
[2] Researchers are actively developing unified frameworks that integrate multiple interpretability methods, with the aim of providing a comprehensive conceptual toolkit for understanding the outputs of complex ML models [30, 31, 39, 60].

Authors' addresses: Andrew Smart, Google Research, San Francisco, USA, andrewsmart@google.com; Atoosa Kasirzadeh, Google Research, San Francisco, USA, atoosa.kasirzadeh@gmail.com.

2024. Manuscript submitted to ACM

should be assessed. A pragmatic and pluralistic approach to interpretability has gained traction, viewing explanations as context-dependent responses to why-questions [12, 31, 42, 43]. On this pluralistic approach, the adequacy of an explanation depends on the specific inquiry.

For simple classification tasks, techniques like saliency maps or feature importance may suffice. For instance, if a model is differentiating between images of cats and dogs, saliency maps could highlight the pixels most influential in the decision-making process. However, for complex and socially-embedded topics — such as biased healthcare algorithms — these model-centric explanations can fall short. Consider an algorithm that predicts hospital readmission risk but systematically underestimates it for certain racial groups. A model-centric explanation might highlight "total healthcare costs incurred in the past year" as an important feature. However, this alone might not fully reveal why the algorithm underestimates risk for a specific racial group. The algorithmic choice could come from the fact that this racial group, due to systemic inequities, have historically been unable to afford adequate healthcare and thus incurred lower costs. As a result, the low value for the "total healthcare costs incurred in the past year" feature does not necessarily indicate better health. Instead, it may suggest unmet healthcare needs, leading to higher readmission rates that the algorithm does not effectively account for. In such cases, interpretations that consider both model-specific details like feature importance and relevant social and structural factors like healthcare affordability disparities among racial groups are crucial for understanding ML predictions or decisions.

In this paper, we draw on social philosophy [17, 25, 26, 67, 68] to advocate for a more comprehensive approach to ML interpretability research, expanding beyond model-centric explanations. We propose incorporating relevant socio-structural explanations to achieve a deeper understanding of ML outputs in domains with substantial societal impact. In the rest of the paper, we introduce the concept of socio-structural explanations and discuss their relevance to understanding ML outputs. We then examine how these explanations can enhance the interpretation of automated decision-making by ML systems in healthcare [49]. Our paper expands the discourse on transparency in machine learning, arguing that it extends beyond model interpretability. We propose that in high-stake decision domains, a socio-structural analysis could be necessary to understand system outputs, uncover societal biases, ensure accountability, and guide policy decisions.

2 INTERPRETABLE ML AND ITS DISCONTENTS

ML interpretability aims to generate human-understandable explanations for model predictions. This process requires the specification of two key components: the explanandum (the phenomenon requiring explanation) and the explanans (the elements providing the explanation). The model's prediction (or decision) typically serves as the explanandum, while visualizations or linguistic descriptions generated via interpretability techniques act as the explanans. To better understand the landscape of interpretability methods, we provide a broad classification of prominent approaches.[3]

Model-centric interpretability approaches can be classified according to various criteria, with one fundamental distinction being between intrinsic and post-hoc interpretability [44]. Intrinsic interpretability achieves transparency by restricting the complexity of the ML model itself, using approaches such as short decision trees or rule-based systems. In contrast, post-hoc interpretability involves applying methods after model training. These methods include SHAP (SHapley Additive exPlanations) values [39], LIME (Local Interpretable Model-agnostic Explanations) [56], saliency maps for neural networks [1], and mechanistic interpretability tools [6].[4] Another popular classification criterion

[3] The interpretable ML literature has grown extensively, making a comprehensive survey beyond the scope of this paper. For recent overviews, see [6, 14, 54, 59].

[4] Post-hoc methods can also be applied to intrinsically interpretable models, such as computing permutation feature importance for decision trees, which can provide additional insights into their decision-making process.

categorizes ML interpretability techniques into two main types: local and global. This categorization offers a complementary perspective by focusing on the scope and depth of the explanations they provide.

Local explanations focus on explaining individual (or a specific group of) predictions or decisions made by a model. Local explanations often use techniques like feature attribution [39, 56] or counterfactual instances [21, 47]. For example, for an image classification model that predicts "dog," a pixel attribution method might highlight the pixels around the dog's ears and tail as being most influential in the prediction "dog." The explanation could be "The model classified this image as a dog primarily because of the distinctive shapes in these highlighted areas (pointing to highlighted pixels in a visualization). The pointed ears here and the curved tail shape here were the most influential features in making this prediction. Other parts of the image, such as the background or the dog's body, had less impact on the classification." For a loan approval ML model, a counterfactual explanation could be "If your income was 5,000 US dollars higher, your loan would have been approved."

Global explanations shed light on the average behavior of the model and provide an overall understanding of how a model works across possible inputs. These methods are often expressed as expected values based on the distribution of the data. Global explanations aim to answer questions like "What features are generally most important for this model's predictions?" or "How does the model behave across different types of inputs?" Techniques for global explanations include partial dependence plots [18] and accumulated local effects [4]. For example, a partial dependence plot, a type of feature effect plot, can show the expected prediction when all other features are marginalized out. In a house price prediction model, a partial dependence plot might show how the predicted price changes as the house size increases, averaged across all other features like location, number of bedrooms, or age of the house. Since global interpretability methods describe average behavior, they are particularly useful when the modeler wants to understand the general mechanisms in the data, debug a model, or gain insights into its overall performance across various scenarios.

Mechanistic interpretations expand upon both local and global explanations. These interpretability tools seek to understand the internals of a model. In the case of neural networks, mechanistic interpretability tools reverse engineer the algorithms implemented by neural networks into concepts, often by examining the weights and activations of neural networks. This approach includes methods such as circuit analysis or dictionary learning for identifying specific subnetworks of neurons within larger models to understand the implementation of particular behavior [16, 63].[5] Mechanistic interpretability is an emerging and highly active area of research, with rapid developments in its neural analysis techniques.

Each of the above-mentioned approaches offers different perspectives on model behavior, ranging from specific instance explanations to overarching principles of operation and fundamental computational mechanisms. The choice of method depends on the specific interpretability goals and the nature of the model being analyzed. There are several acknowledged limitations to existing interpretability approaches.

First, interpretability techniques can be brittle, sensitive to the target of interpretation [3, 45, 65], to minor perturbations in model parameters [20] or input data [52]. This fragility raises concerns about the reliability and robustness of generated explanations using interpretability methods, especially in real-world scenarios where models are subject to noisy data and evolving conditions. Recent work on mechanistic interpretability has begun to discover features

[5] Neurons in neural networks can be monosemantic (representing a single concept) or polysemantic (representing multiple unrelated concepts). Monosemantic neurons activate for a single semantic concept, suggesting a one-to-one relationship between neurons and features [63]. However, neurons are often polysemantic, activating for multiple unrelated concepts, complicating network interpretation. For instance, researchers have empirically shown that for a certain language model, a single neuron can correspond to a mixture of academic citations, English dialogue, HTTP requests, and Korean text [9]. Polysemanticity makes it difficult to reason about the behavior of the network in terms of the activity of individual neurons.

of large language models that are more robust [50, 63, 64]. However, there is still significant progress to be made in developing consistently reliable interpretability methods [65].

Second, for a given model and input, there may be multiple valid explanations, each highlighting different aspects of the prediction-making or decision-making process [62]. This multiplicity embodies both a feature and a bug: as a feature, it reflects the need for a multi-faceted understanding of the complexity of ML predictions; as a bug, it introduces potential confusion and conflicting interpretations, challenging efforts to identify the most relevant or meaningful explanation. Consider, for instance, an ML model used in hiring decisions that recommends not to hire a particular candidate. A feature importance analysis might indicate that the candidate's educational background was the primary factor in the decision. However, a counterfactual explanation might suggest that changing the candidate's gender would alter the outcome. Simultaneously, a SHAP analysis could show that a combination of factors including work experience, interview performance, and age contributed to the decision. Each of these explanations provides insight into the model's reasoning, but emphasizes different aspects, some of which may be more socially sensitive or legally problematic than others. This diversity of explanations challenges practitioners in determining which aspects are most crucial for the model's behavior. Moreover, different stakeholders - such as job applicants, hiring managers, and legal compliance officers - might prefer or trust certain types of explanations over others, further complicating the practical application of these interpretability methods [7]. Despite the growing number of interpretability approaches, there is a lack of standardized benchmarks and evaluation frameworks to assess their legal and ethical relevance and compare their performance [2, 29].

Third, we still have no provable guarantee that a post-hoc explanation accurately reflects the true reasoning behind a model's prediction [1, 58, 65]. Explanations may be overly simplistic, highlight irrelevant features, or even be misleading, potentially leading to incorrect conclusions about the model's behavior. The potential lack of faithfulness is particularly problematic in high-stakes domains where decisions have significant consequences. For example, a counterfactual explanation for a loan denial might suggest that increasing income would lead to approval. However, the true cause might be a complex interaction of credit history and debt-to-income ratio, not captured by the explanation [34, 62]. Given these limitations, researchers are exploring novel methods to enhance our understanding of ML models and their predictions (or decisions).

A promising approach to enhance model transparency involves expanding the scope of interpretations beyond the internal mechanics of the model itself. This expanded perspective recognizes that ML models do not operate in isolation, but within complex social and institutional contexts that can significantly influence their behavior and impact. Here, we propose a new perspective to interpreting ML outputs that incorporates relevant social and structural factors into transparency demands. In particular, we think that in certain situations, the soundness and stability of ML explanations can be improved by appealing to what we call socio-structural explanations that are *external* to an ML model. Our thesis is that in some socially salient applications of ML models, perhaps the most important constraints on model behavior are *external* to the model itself. Extending the idea that the machine learner is not only the inert model, but includes the human developers, uses and surrounding social relations and practices [55], we propose to explain the behavior of a model, in such instances, *given* its place in a *social structure*. We call such explanations *socio-structural* explanations. In order to understand socio-structural explanations, we first need to know what are social structures?

3 SOCIAL STRUCTURES AND SOCIO-STRUCTURAL EXPLANATIONS

Social structures are the underlying realities that shape our social lives, influencing our choices, opportunities, and experiences [25, 67]. They are the invisible scaffolding of society, both constraining and enabling our individual and collective actions. They give rise to social hierarchies through institutions, policies, economic systems, and cultural or normative belief systems such as race or socioeconomic status [8]. Social structures manifest in various forms, from the subtle influence of societal norms to the explicit impact of legal frameworks, creating a multilayered reality that shapes our experiences and opportunities.

Social and political philosopher, Iris Marion Young [67, 68], defines social structures as the interplay of institutional rules, interactive routines, resource mobilization, and physical infrastructure. These enduring elements shape the context within which individuals act, offering both opportunities and limitations.[6] These structures, while socially constructed, possess a reality for exerting tangible influences on individuals and institutions [24, 68]. They are powerful forces that can constrain and enable actions, cause the specific distribution of resources, and define social roles and expectations. Social structures can explain persistent patterns and circumstances in society, such as racial inequality or gender disparities. To get more concrete, let us analyze the notion of social structures in the context of a socio-structural explanation, borrowing a simple example from Garfinkel [19]:

> Suppose that, in a class I am teaching, I announce that the course will be “graded on a curve,” that is, that I have decided beforehand what the overall distribution of grades is going to be. Let us say, for the sake of the example, that I decide that there will be one A, 24 Bs, and 25 Cs. The finals come in, and let us say Mary gets the A. She wrote an original and thoughtful final.

Garfinkel [19] argues that the explanation "She wrote an original and thoughtful final" is inadequate to answer the explanation-seeking question "Why did Mary get an A?" In a curved grading system, achieving the sole A grade requires more than just quality work. For Mary to earn the only A in the class, her final would need to be the best. If the instructor had not implemented a grading curve, multiple students could have earned As by producing thoughtful and original finals. Garfinkel elaborates on this point, stating "So it is more accurate to answer the question by pointing to the relative fact that Mary wrote the best paper in the class" [19, p. 41]. Mary’s A grade was not solely a result of her individual performance, but also a consequence of her relative standing among peers, combined with the specific grading structure that emphasized this comparative aspect. The grading structure, in this case, serves as a crucial contextual element shaping the explanation. More precisely, the structural aspect of this explanation is "the predetermined grading curve that limited the number of As to one," while the social aspect is "Mary’s performance relative to her peers (she wrote the best paper in the class)."

Here is a different example for further clarification of the notion of socio-structural explanation. Consider the following explanatory question: “Why do women continue to be economically disadvantaged relative to men (as opposed to reaching economic parity with men?)” [25]. Haslanger [25] argues that we can have (at least) three explanations for this question: biological, individualistic, and structural.

Biologistic explanation: Women are inherently less capable than men in biological qualities deemed necessary (such as intelligence or competitiveness) for success in high-paying jobs.

[6] Social structures, according to Young [67, p.111], are defined as follows:

> As I understand the concept, the confluence of institutional rules and interactive routines, mobilization of resources, as well as physical structures such as buildings and roads. These constitute the historical givens in relation to which individuals act, and which are relatively stable over time. Social structures serve as background conditions for individual actions by presenting actors with options; they provide “channels” that both enable action and constrain it.

Individualistic explanation: Women, to a greater extent than men, prioritize child-rearing over pursuing high-paying careers, thus voluntarily sacrificing economic success for the perceived rewards and satisfactions of motherhood.

Structural explanation: Women are embedded within a self-reinforcing economic structure that systematically disadvantages them through institutional practices, social norms, and power dynamics.

Each of these explanations refers to different causes, operating at distinct levels of analysis. The biologistic and individualistic explanations focus on factors intrinsic to individuals or groups, without considering the broader socio-structural context. In contrast, the structural explanation situates individual actions and outcomes within a larger system of interconnected social forces. If the social structure is in place, then we can view individuals as occupying specific "nodes" within a complex social network or structure. The socio-structural explanation posits that gender wage disparities arise from the complex interplay of societal, economic, and institutional factors that collectively shape opportunities and constraints. Given the socio-structural limitations in place, we can explain why women, at the population level, experience economic disadvantages compared to men based on their position within the social structure. In this context, "social structure" refers to the complex network of institutions, relationships, and cultural norms that organize society. It includes economic systems that historically undervalue work traditionally performed by women, political institutions that may underrepresent women's interests, and educational structures that can reinforce gender stereotypes. Additionally, it includes cultural norms that influence career choices and work-life balance expectations, organizational hierarchies that often favor male leadership, and legal frameworks that may inadequately address gender discrimination.

Women's place within this multifaceted social structure often results in reduced access to resources, limited decision-making power, and fewer opportunities for advancement, collectively contributing to persistent economic disparities at the population level. The socio-structural approach to explanation, when rigorously applied, offers valuable insights. It demonstrates how individual choices and actions can be profoundly shaped by the surrounding social structures. By highlighting the influence of broader structural forces on seemingly personal decisions, it reveals patterns often operating beyond an individual's immediate awareness or control.

Let us draw a close analogy between the above instances of socio-structural explanations and a toy example of interpreting an ML model's output. Consider an ML-powered hiring model that consistently recommends male candidates over female candidates for senior executive positions in a tech company. An initial explanation of the recommendations generated by a SHAP interpretability method might say: "The model recommends male X over female Y because X's features contribute more positively to the model's output. Specifically, X's 10 years of tech leadership experience contributes +0.4 to the score, while Y's 7 years contributes only +0.2." Let us assume similar explanations (relating years of tech leadership experience to the recommendation score) are generated for a population of females. These explanations might fail to capture the full picture.

Upon deeper investigation, an auditor team uncovers a more complex and nuanced reality. First, the auditors find that the ML model was trained on the company's historical hiring data from 2000-2020, which included 85% male executives. This data reflects the company's past hiring practices, which favored men for leadership roles. The socio-structural aspect here is the historical underrepresentation of women in executive positions, rooted in long-standing societal norms and institutional practices. A socio-structural explanation could look like: "The model's bias reflects decades of systemic exclusion of women from leadership roles in the tech industry, perpetuating a cycle where the lack of female representation in executive positions reinforces the perception that these roles are best suited for men." Second, the ML model places high importance on continuous work experience, with any gap longer than 6 months reducing a candidate's score by 0.1 per year. 40% of female candidates had career gaps averaging 2.5 years, compared to 10% of male candidates averaging 1 year, often coinciding with childbearing ages. This reflects the socio-structural reality of women bearing a disproportionate responsibility for child-rearing and family care, leading to more frequent and longer career interruptions. A socio-structural explanation could look like: "The model's penalty for career gaps disproportionately impacts women due to societal expectations and norms that place the primary burden of childcare and family responsibilities on women, resulting in more frequent and longer career interruptions that are then interpreted by the model as reduced qualifications." Third, the model does not consider geographic location in its evaluation. However, geographic disparities affect job availability and commute times, disproportionately impacting women with childbearing responsibilities. This reflects socio-structural expectations around family care that often limit women's job options to those closer to home or with flexible hours. A socio-structural explanation could look like: "The model's failure to account for geographic factors overlooks the societal expectations that often constrain women's job choices based on proximity to home and flexibility for family care. This oversight particularly disadvantages women who may be highly qualified but limited in their job options due to these socially imposed constraints."

Producing rigorous socio-structural explanations can be challenging as it requires significant sociological understanding and interdisciplinary expertise. However, once obtained, these explanations enable novel forms of interventions. Here are some examples of possible interventions enabled by obtaining socio-structural explanations. The first is to modify the model to cap the maximum contribution of "continuous experience" at +0.2. The second is to introduce a new feature "diverse experience" that values varied career paths, including those with gaps. The third is to augment the training data with 500 profiles of successful executives who have had career gaps of 1-3 years, ensuring at least 50% are women. The fourth is to implement a company-wide policy requiring human review for any candidate the ML system ranks lower primarily due to career gaps (>0.2 score difference). This toy example is supposed to highlight that integrating socio-structural explanations into the ML transparency toolkit enables us to transcend superficial model-centric solutions (when relevant) and address the fundamental causes underlying ML outputs.

Of course, the specific interventions depend on what we want to change and the particular context of the problem at hand. Socio-structural explanations are not always useful or applicable in every situation. The effectiveness of these explanations and subsequent interventions can vary based on the complexity of the social systems involved, the quality of available data, and the specific goals of the analysis. In some cases, other approaches might be more appropriate or effective.

In the following section, we examine a case study of algorithmic deployment in healthcare decision-making, highlighting the critical relevance of socio-structural explanations in this context. This analysis demonstrates how a deeper understanding of social structures can inform more effective strategies for developing and implementing algorithmic systems in high-stakes decision domains. While our original focus was on socio-structural explanations for ML systems, we recognize that the importance of these explanations generalizes to a broader range of automated decision systems.

4 SOCIO-STRUCTURAL EXPLANATIONS OF RACIAL BIAS IN HEALTH-CARE ALGORITHMS

A widely discussed example in the growing body of literature on algorithmic bias is the study by Obermeyer et al. [49]. This research revealed that a commonly used US hospital predictive algorithm for allocating scarce healthcare resources systematically discriminated against Black patients. Specifically, the algorithm assigned lower risk scores to Black patients who were equally in need as their White counterparts. The root cause was the algorithm's use of healthcare costs as a proxy for "healthcare need" [35]. This approach led to a significant underestimation of health risks for Black patients who, on average, incurred lower healthcare costs than White patients with similar chronic conditions due to systemic disparities in care access and quality [49].

Empirical investigations demonstrated that the care provided to Black patients cost an average of USD 1,800 less per year than the care given to a white person with the same number of chronic health problems. At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses [49]. The algorithm predicted that this disparity in spending corresponded to a similar disparity in actual health-care needs and therefore risk score. Consequently, Black people had to be much sicker in order to be referred for treatment or other resources. The algorithm's prediction of health needs is, in fact, a prediction on health costs [49].

When algorithmic decision systems fail in consequential domains like health-care, the repercussions can be severe, potentially leading to patient deaths. It is crucial to understand the reasons and causes for such failures. Therefore, explaining the "why" behind these failures through the analysis of failed outputs is critical. One prevalent type of failure is algorithmic bias that perpetuates existing socio-structural inequalities, such as structural racism [5, 23, 49, 53]. Structural racism refers to the complex ways in which historical and contemporary racial inequities are reproduced through interconnected societal systems like healthcare, education, housing, and the criminal justice system [51]. Even when race is not explicitly considered, its influence can be deeply embedded in the data, shaping associations and outcomes [57]. In the context of this instance, the following explanatory question demands a response: Why did this algorithm systematically discriminate against Black people?

To answer this question, we must consider both the interpretation in reference to the model and the broader socio-structural context in which it operates.[7] Obermeyer et al. [49] show that this particular algorithm discriminated against Black patients due to its use of healthcare costs as a proxy for "healthcare need." This choice reflects a fundamental misunderstanding of the relationship between costs and needs in a healthcare system marked by systemic racial disparities. Obermeyer et al. [49] demonstrated that conditioning on healthcare costs is the mechanism by which the bias arises in this case, and we must change the data we feed the algorithm and use new labels that better reflect social reality, which in turn requires deep understanding of the domain, the ability to identify and extract relevant data elements, and the capacity to iterate and experiment [49]. The socio-structural interpretation of the algorithm's behavior is as follows.

> The algorithm discriminates against Black patients because it is designed and deployed in a healthcare system characterized by longstanding racial inequities. By encoding healthcare costs as a proxy for healthcare needs, the algorithm inadvertently encodes and perpetuates systemic disparities in care access and quality. Black patients, on average, incur lower healthcare costs not because they are healthier, but due to historical patterns of exclusion, lack of access to care, and underinvestment in healthcare resources for Black communities. The algorithm interprets these lower costs as lower needs, thereby underestimating the health risks for Black patients and perpetuating a cycle of inadequate care allocation. This reflects how the algorithm manages to reproduce structural racism through its uncritical use of data that embodies these systemic inequalities.

It remains challenging for practitioners to identify the harmful repercussions of their own systems prior to deployment, and, once deployed, emergent issues can become difficult or impossible to trace back to their source [53].

[7] For a discussion of the levels of interpretation see Creel [12] and Kasirzadeh and Klein [33].

Unfortunately, many failures of algorithmic decision systems in the healthcare industry disproportionately impact people or communities who have been put already in a structurally vulnerable social positions [36]. This can be due to many factors. However, a consistent theme in the study of these failures, that is often only revealed after the fact, is that there is a lack of socio-structural understanding among the designers and users of these systems [40]. The study presented in this section exemplifies this challenge. Employing model-centric explanations would likely highlight the importance of the cost feature to algorithmic output, but would not expose the underlying racial bias originating from historical and systemic inequalities in healthcare access and delivery. In this context, socio-structural explanations consider the relevant societal context in which the model operates, in relation to relevant historical biases, societal norms, and institutional practices.

5 IMPLICATIONS FOR ML TRANSPARENCY RESEARCH AND CONCLUSION

ML research and practice are fundamentally shaped by the approaches adopted by practitioners. These approaches influence the entire process: from the questions asked and data collected, to the choice of objective functions and the selection of proxy or target variables for optimization. Throughout this paper, we have argued that model-centric explanations, while valuable, can be inadequate for comprehensively understanding whether a model truly benefits or potentially harms people. This inadequacy is particularly pronounced in high-stakes domains where ML models are often developed and deployed into complex social and structural contexts without sufficient domain-specific theoretical understanding. We have argued that to meaningfully interpret the *social* predictions (or decisions) of models in high-stake domains, a deep socio-structural understanding is required.

One challenge lies in that many ML practitioners and researchers may not feel adequately equipped to analyze and respond to social structures. Alternatively, they may be hindered from leveraging social structural knowledge due to constraints in time, training, incentives, or resources [61]. This gap between technical expertise and socio-structural understanding presents a significant hurdle in developing truly beneficial ML systems. Algorithmic transparency and accountability research in ML is often motivated by the need to foster trust in these systems. Much of this research rightly argues for the critical importance of model-centric interpretations [20, 22, 39]. However, the demands for transparency of ML models must extend beyond model-centric details to encompass socio-structural factors in socially-salient prediction or decision domains. Producing new, more representative labels and objectives for ML models requires a deep understanding of the domain, the ability to identify and extract relevant data elements, and the capacity to iterate and experiment [49].

The importance of socio-structures has been increasingly recognized in recent literature on algorithmic justice [27, 32, 37]. These works argue for a more holistic approach to ML development and deployment, one that considers not just model-centric measures but also societal impacts. In light of these considerations, we call for further research into the integration of socio-structural understanding into different stages of the ML lifecycle, from problem formulation and data collection to model development, deployment, and ongoing monitoring. We think that sometimes socio-structural interpretations can reveal causally-relevant reasons for why an algorithm behave in a certain way for a certain population. By doing so, we can work towards ML systems that are not only technically proficient but also socially aware and beneficially aligned.

6 ACKNOWLEDGMENTS

We would like to thank Donald Martin, Been Kim, Darlene Neal, Mayo Clinic Accelerate Program and The Impact Lab team at Google Research for helpful discussions and feedback on earlier drafts of this paper.

REFERENCES

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems* 31 (2018).

[2] Chirag Agarwal, Eshika Saxena, Satyapriya Krishna, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. 2022. OpenXAI: Towards a Transparent Evaluation of Model Explanations. *arXiv preprint arXiv:2206.11104* (2022).

[3] David Alvarez-Melis and Tommi S Jaakkola. 2018. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049* (2018).

[4] Daniel W Apley and Jingyu Zhu. 2020. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82, 4 (2020), 1059–1086.

[5] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[6] Leonard Bereska and Efstratios Gavves. 2024. Mechanistic Interpretability for AI Safety–A Review. *arXiv preprint arXiv:2404.14082* (2024).

[7] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 648–657.

[8] Philippe Bourgois, Seth M Holmes, Kim Sue, and James Quesada. 2017. Structural vulnerability: operationalizing the concept to address health disparities in clinical care. *Academic medicine: journal of the Association of American Medical Colleges* 92, 3 (2017), 299.

[9] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. 2023. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits* (2023). https://transformer-circuits.pub/2023/monosemantic-features

[10] Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. 2019. Activation atlas. *Distill* 4, 3 (2019).

[11] Valerie Chen, Jeffrey Li, Joon Sik Kim, Gregory Plumb, and Ameet Talwalkar. 2022. Interpretable Machine Learning: Moving from mythos to diagnostics. *Queue* 19, 6 (2022), 28–56.

[12] Kathleen A Creel. 2020. Transparency in complex computational systems. *Philosophy of Science* 87, 4 (2020), 568–589.

[13] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[14] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. 2023. Explainable AI (XAI): Core ideas, techniques, and solutions. *Comput. Surveys* 55, 9 (2023), 1–33.

[15] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652* (2022).

[16] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A Mathematical Framework for Transformer Circuits. https://transformer-circuits.pub/2021/framework/index.html.

[17] Brian Epstein. 2015. *The ant trap: Rebuilding the foundations of the social sciences*. Oxford University Press.

[18] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[19] Alan Garfinkel. 1981. *Forms of Explanation: Rethinking the Questions in Social Theory*. Yale University Press, New Haven.

[20] Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 3681–3688.

[21] Riccardo Guidotti. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery* (2022), 1–55.

[22] Tessa Han, Suraj Srinivas, and Himabindu Lakkaraju. 2022. Which Explanation Should I Choose? A Function Approximation Perspective to Characterizing Post hoc Explanations. *arXiv preprint arXiv:2206.01254* (2022).

[23] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 501–512.

[24] Sally Haslanger. 2012. *Resisting reality: Social construction and social critique*. Oxford University Press.

[25] Sally Haslanger. 2016. What is a (social) structural explanation? *Philosophical Studies* 173, 1 (2016), 113–130.

[26] Sally Haslanger. 2020. Failures of methodological individualism: the materiality of social systems. (2020).

[27] Anna Lauren Hoffmann. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22, 7 (2019), 900–915.

[28] Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. 2022. Explainable AI methods-a brief overview. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*. Springer, 13–38.

[29] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems* 32 (2019).

[30] Tobias Huber, Katharina Weitz, Elisabeth André, and Ofra Amir. 2021. Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps. *Artificial Intelligence* 301 (2021), 103571.

[31] Atoosa Kasirzadeh. 2021. Reasons, values, stakeholders: A philosophical framework for explainable artificial intelligence. *arXiv preprint arXiv:2103.00752* (2021).

[32] Atoosa Kasirzadeh. 2022. Algorithmic fairness and structural injustice: Insights from feminist political philosophy. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 349–356.

[33] Atoosa Kasirzadeh and Colin Klein. 2021. The ethical gravity thesis: Marrian levels and the persistence of bias in automated decision-making systems. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 618–626.

[34] Atoosa Kasirzadeh and Andrew Smart. 2021. The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 228–236.

[35] Heidi Ledford. 2019. Millions of black people affected by racial bias in health-care algorithms. *Nature* 574, 7780 (2019), 608–610.

[36] David Leslie, Anjali Mazumder, Aidan Peppin, Maria K Wolters, and Alexa Hagerty. 2021. Does "AI" stand for augmenting inequality in the era of covid-19 healthcare? *bmj* 372 (2021).

[37] Ting-An Lin and Po-Hsuan Cameron Chen. 2022. Artificial intelligence in a structurally unjust society. *Feminist Philosophy Quarterly* 8, 3/4 (2022).

[38] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[39] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[40] Donald Martin Jr, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S Isaac. 2020. Participatory problem formulation for fairer machine learning through community based system dynamics. *arXiv preprint arXiv:2005.07572* (2020).

[41] Nijat Mehdiyev and Peter Fettke. 2021. Explainable artificial intelligence for process mining: A general overview and application of a novel local explanation approach for predictive process monitoring. *Interpretable Artificial Intelligence: A Perspective of Granular Computing* (2021), 1–28.

[42] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547* (2017).

[43] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. 279–288.

[44] Christoph Molnar. 2020. *Interpretable machine learning*. Leanpub.

[45] Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. 2020. General pitfalls of model-agnostic interpretation methods for machine learning models. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*. Springer, 39–68.

[46] Andrea Morichetta, Pedro Casas, and Marco Mellia. 2019. EXPLAIN-IT: Towards explainable AI for unsupervised network traffic analysis. In *Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks*. 22–28.

[47] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 607–617.

[48] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217* (2023).

[49] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[50] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895* (2022).

[51] Kristen Pallok, Fernando De Maio, and David A Ansell. 2019. Structural racism—a 60-year-old black woman with breast cancer. *New England Journal of Medicine* 380, 16 (2019), 1489–1493.

[52] Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. 2019. Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems* 32 (2019).

[53] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.

[54] Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. 2023. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *2023 ieee conference on secure and trustworthy machine learning (satml)*. IEEE, 464–483.

[55] Tyler Reigeluth and Michael Castelle. 2021. What Kind of Learning Is Machine Learning? In *The Cultural Life of Machine Learning*. Springer, 79–115.

[56] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[57] Whitney R Robinson, Audrey Renson, and Ashley I Naimi. 2020. Teaching yourself about structural racism will improve your machine learning. *Biostatistics* 21, 2 (2020), 339–344.

[58] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.

[59] Waddah Saeed and Christian Omlin. 2023. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems* 263 (2023), 110273.

[60] Jessica Schrouff, Sebastien Baur, Shaobo Hou, Diana Mincu, Eric Loreaux, Ralph Blanes, James Wexler, Alan Karthikesalingam, and Been Kim. 2021. Best of both worlds: local and global explanations with human-understandable concepts. *arXiv preprint arXiv:2106.08641* (2021).

[61] Shreya Shankar, Rolando Garcia, Joseph M Hellerstein, and Aditya G Parameswaran. 2022. Operationalizing Machine Learning: An Interview Study. *arXiv preprint arXiv:2209.09125* (2022).

[62] Emily Sullivan and Atoosa Kasirzadeh. 2024. Explanation Hacking: The perils of algorithmic recourse. *arXiv preprint arXiv:2406.11843* (2024).

[63] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread* (2024). https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

[64] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593* (2022).

[65] David S Watson. 2022. Conceptual challenges for interpretable machine learning. *Synthese* 200, 2 (2022), 65.

[66] David S Watson, Limor Gultchin, Ankur Taly, and Luciano Floridi. 2021. Local explanations via necessity and sufficiency: unifying theory and practice. In *Uncertainty in Artificial Intelligence*. PMLR, 1382–1392.

[67] Iris Marion Young. 2006. Responsibility and global justice: A social connection model. *Social philosophy and policy* 23, 1 (2006), 102–130.

[68] Iris Marion Young. 2010. *Responsibility for justice*. Oxford University Press.