CAMBRIDGE
UNIVERSITY PRESS

TRANSLATIONAL ARTICLE

Explainable machine learning for public policy: Use cases, gaps, and research directions

Kasun Amarasinghe[1,2] , Kit T. Rodolfa[1,2] , Hemank Lamba[1,2] and Rayid Ghani[1,2,*]

[1]Machine Learning Department, Carnegie Mellon University, 4902 Forbes Avenue, Pittsburgh, Pennsylvania, 15213, USA
[2]Heinz College of Information Systems and Public Policy, Carnegie Mellon University, 4902 Forbes Avenue, Pittsburgh, Pennsylvania, 15213, USA
*Corresponding author. E-mail: rayid@cmu.edu

Abstract

Explainability is highly desired in machine learning (ML) systems supporting high-stakes policy decisions in areas such as health, criminal justice, education, and employment. While the field of explainable ML has expanded in recent years, much of this work has not taken real-world needs into account. A majority of proposed methods are designed with *generic* explainability goals without well-defined use cases or intended end users and evaluated on simplified tasks, benchmark problems/datasets, or with proxy users (e.g., Amazon Mechanical Turk). We argue that these simplified evaluation settings do not capture the nuances and complexities of real-world applications. As a result, the applicability and effectiveness of this large body of theoretical and methodological work in real-world applications are unclear. In this work, we take steps toward addressing this gap for the domain of public policy. First, we identify the primary use cases of explainable ML within public policy problems. For each use case, we define the end users of explanations and the specific goals the explanations have to fulfill. Finally, we map existing work in explainable ML to these use cases, identify gaps in established capabilities, and propose research directions to fill those gaps to have a practical societal impact through ML. The contribution is (a) a methodology for explainable ML researchers to identify use cases and develop methods targeted at them and (b) using that methodology for the domain of public policy and giving an example for the researchers on developing explainable ML methods that result in real-world impact.

Policy Significance Statement

Despite a rich body of methodological work in explainable ML, little guidance exists for building systems that meet the needs of actual policy applications. This article seeks to fill that void by mapping out explainability use cases in public policy settings and comparing the capabilities of existing methods against the requirements of each use case's stakeholders. We believe that this work serves public policy in two ways: (a) for researchers, a call for empirical, application-focused development and evaluation of explainability methods that will lead to systems better suited to provide social impact; and (b) for policymakers and ML practitioners, a guide to navigating the complex landscape of ML explainability when designing and evaluating applied ML systems that support their policy objectives.

CrossMark

1. Introduction

Machine learning (ML) systems are increasingly supporting high-stakes public policy decisions in areas such as criminal justice, education, healthcare, and social services (Caruana et al., 2015; Bauman et al., 2018; Ye et al., 2019; Potash et al., 2020; Rodolfa et al., 2020). As users of these systems have grown beyond ML experts and the research community, the need to better interpret and understand them has grown as well, particularly in the context of high-stakes decisions that affect individuals' health or well-being (Lakkaraju et al., 2016; Lipton, 2018; Rudin, 2019). Likewise, new legal frameworks reflecting these needs are beginning to emerge, such as the *right to explanation* in the European Union's General Data Protection Regulation (Goodman and Flaxman, 2017).

Against this background, research into *explainability/interpretability*[1] of ML models has experienced rapid expansion and innovation in recent years, with a focus on method development. A range of methods have been developed that broadly fall into two categories: (a) inherently interpretable models (Ustun et al., 2013; Caruana et al., 2015; Lakkaraju et al., 2016; Yang et al., 2017; Rudin, 2019) and (b) post hoc methods for explaining (opaque) complex models and/or their predictions (Bach et al., 2015; Ribeiro et al., 2016, 2018; Lundberg and Lee, 2017; Lundberg et al., 2018a, 2018b; Wachter et al., 2018; Mothilal et al., 2020). While this expansion of the field has yielded a rich body of methodological work, recently, the community has begun to highlight the shortfalls such as the lack of consistent language and definitions, the lack of clearly defined explainability goals, and desiderata; and the lack of consensus on metrics and methods of evaluating the quality of explanations (Doshi-Velez and Kim, 2017; Lipton, 2018; Weller, 2019; Buçinca et al., 2020; Hase and Bansal, 2020; Sokol and Flach, 2020; Bhatt et al., 2020a, 2020b; Chen et al., 2022). In addition to the critique above, we argue that there are two key areas where most existing work related to explainable ML methods falls short:

1. Explainability methods are often developed as "general-purpose" methods with a broad and loosely defined goal, such as perceived transparency, and not to address specific needs of real-world use cases.
2. Explainability methods are not rigorously evaluated to adequately reflect their effectiveness in real-world settings. Barring a few exceptions (Caruana et al., 2015; Lundberg et al., 2018b; Ustun et al., 2019; Jesus et al., 2021), much of the existing work is designed and developed for benchmark classification problems, often with synthetic data and usually validated with user studies limited to users in research settings such as Amazon Mechanical Turk (AMT; Simonyan et al., 2013; Zeiler and Fergus, 2014; Bach et al., 2015; Ribeiro et al., 2016; Lundberg and Lee, 2017; Plumb et al., 2018; Hu et al., 2019).

The result is a body of methodological work without clearly identified use cases and, more importantly, without established real-world utility, making it difficult for practitioners to select and deploy these methods with any confidence. A necessary first step for filling these gaps is clearly defining how explainable ML fits into a decision-making process. As explainability is not a monolithic concept and can play different roles in different applications (Lipton, 2018; Molnar, 2019), this process requires extensive domain/application-specific efforts.

In this article, we use our experience working with government agencies and nonprofits and focus on applications of ML to public policy problems. Among the broad range of intervention points that the domain of policy presents to ML (e.g., policy design, evaluation, and administration), we focus our attention on policy administration tasks where predictive ML models are used to support human decisions with objectives of improving the efficiency of resource usage, the effectiveness of interventions, and equity of outcomes. We seek to define the role of explainable ML in these domains and how we can use it to improve policy and social outcomes. To that end, this article has the following contributions:

---

[1] We combine the two terms *interpretability* and *explainability* and use both terms to refer to the ability to understand, interpret, and explain ML models and their predictions.

1. identifying the primary use cases of ML explanations in public policy applications;
2. for each use case, identifying the goals of the explanation methods, the end users, and the explanation needs;
3. identifying research gaps by comparing the known capabilities of the existing body of work to the needs of the use cases;
4. proposing research directions to develop effective explainable ML systems that would be targeted for the needs of real-world use cases and lead to improved policy decisions and consequently improved societal outcomes.

The primary goal of this work is to bridge the gap between methodological research in explainable ML and the needs of policy applications. We believe that effective human–ML collaborative decision-making can profoundly impact policy decision-making processes, and the explainability of ML systems plays a critical role in human–ML collaboration. Thus, bridging the gap between explainable ML methods and applications is paramount. As computer scientists who develop and apply ML algorithms to improve policy decision-making processes, this article is our attempt at connecting the ML research community with problems in public policy where explainable ML can impact consequential decisions. It is worth noting that there have been other pieces in the literature that are similarly motivated in understanding how we can use explainable ML in practical applications (Hong et al., 2020; Bhatt et al., 2020a; Belle and Papantonis, 2021). These efforts have primarily focused on identifying how the ML community is using existing interpretability methods. For instance, Bhatt et al. (2020a, 2020b) and Hong et al. (2020) conduct semi-structured interviews with different stakeholders in industry to understand how they incorporate interpretable ML in their workflows and gain valuable insights into how ML practitioners perceive the methods of explainable ML. We believe that our work supplements said pieces through an in-depth analysis of a single domain where we look at potential uses, the needs, where there are gaps in current research in meeting those needs, and how the practitioners and researchers can work together in bridging those gaps. We do not intend this work to be a thorough survey of existing work in explainable ML (since there are already excellent articles on that topic; Adadi and Berrada, 2018; Guidotti et al., 2018; Arya et al., 2019; Molnar, 2019; Bhatt et al., 2020b) but rather to highlight the needs of the domain, map the capabilities of existing approaches to those needs, identify gaps, and propose concrete steps to bridge those gaps. The primary audience of this work is the ML research community that designs and develops explainable ML systems that may be implemented in public policy decision-making systems. We believe that this discussion will serve as a framework for designing explainable ML methods and evaluation setups with an understanding of the following:

1. the purpose the explanations serve and the related policy/societal outcome;
2. the end user of the explanations, the exact decisions they would make based on the explanations, and the intended impact of explanations on their decisions;
3. how to measure the effectiveness of generated explanations in helping end users make better decisions that result in improved public outcomes (e.g., metrics that reflect decision outcomes).

Furthermore, we believe that this work could serve as a guide for policy practitioners who procure and embed ML systems in their decision processes to perform more informed evaluations of the explainable ML systems they procure.

Although the focus on public policy applications reflects the area of expertise of the authors and a domain that is beginning to use ML tools for assisting many high-stakes decisions, we believe that our approach to defining the role of explainable ML in this setting will be valuable as a template for other domains as well.

2. Use of Machine Learning in Public Policy Problems

ML models can analyze large amounts of data to identify patterns and make predictions about future events (e.g., the risk of an evicted individual ending up homeless in the next year, the risk of a student not

graduating high school on time, processing legislative bills to understand the policy areas covered in the bill). These predictions can provide data-driven insights to supplement human expertise to inform decision making and the policy domain presents a range of such decision points. However, it is important to note that policy decision-making is a complex, human, and political process, and there are many challenges to using ML in this domain. For instance, policymaking typically involves trade-offs between competing social values (Parkhurst, 2016; Saltelli and Giampietro, 2017), whereas ML algorithms require explicitly defined objectives and weighing competing objectives (Coyle and Weller, 2020). Additionally, while ML predictions can provide useful information, policy decisions are typically not solely based on technical evidence (Parkhurst, 2016), and the power of persuasion and in some cases manipulation, plays a critical role in legislative processes (Zahariadis, 2003; Cairney and Oliver, 2017).

In this work, we focus on policy administration decisions where ML predictions assist resource allocation and intervention decisions at a highly granular level (e.g., predicting the risk of future mental health crises to do individual-level proactive mental health outreach). Typically, these systems are designed to improve the efficiency of resource utilization, intervention effectiveness, and equity of outcomes. To mitigate the ambiguity and uncertainty inherent to policy processes, we assume a continuous partnership between partnering policy practitioners and ML practitioners in defining the goals, parameters of operationalizing ML predictions, how to measure success, possible risks, and mitigation strategies (e.g., bias and equity). In this work, we draw on our experience in partnering with governments to develop and implement human–ML collaborative policy administrative systems where ML predictions (and potential ML explanations) supplement decision-makers' domain expertise.

To illustrate the applicability of ML to policy administration settings, we focus on the common task of early warning systems (EWSs) that are prevalent in different policy domains. In an EWS, the ML model is used to identify entities (e.g., people, schools, buildings, and locations) for some intervention, based on a predicted risk of some (often adverse) outcome, such as an individual getting diagnosed with a disease in the next year, a student not graduating high school or college on time, a tenant getting harassed by their landlord, or for a child getting lead poisoning within the next year (Bauman et al., 2018; Ye et al., 2019; Rodolfa et al., 2020). While there are several other policy problem templates that ML is used for, such as inspection targeting, scheduling, routing, and policy evaluation, we use EWSs to illustrate our ideas in this article.

2.1. Characteristics of ML applications in public policy

Several characteristics of typical public policy problems set them apart from standard benchmark ML problems and datasets often used to evaluate newly proposed algorithms.

2.1.1. Nonstationary environments

In a policy context, ML models use data about historical events to predict the likelihood of either the occurrence of an event in the future or the existence of a present need, and the context around the problem changes over time. This nonstationary nature of the data introduces strong temporal dependencies that should be considered throughout the modeling pipeline and makes these models susceptible to errors such as data leakage (Kaufman et al., 2011; Samala et al., 2020). For instance, the use of standard randomized k-fold cross-validation as a model selection strategy can create training sets with information from the future, which would not have been available at model training time.

2.1.2. Evaluation metrics reflect real-world resource constraints

The mental health outreach in Bauman et al. (2018) was limited by staffing capacity to intervene on only 200 individuals at a time, and the rental inspections team in Ye et al. (2019) could only inspect around 300 buildings per month. Resource constraints such as these are inherent in policy contexts, and the metrics used to evaluate and select models should reflect the deployment context. As such, these applications fall into the *top-k* setting, where the task involves selecting exactly k instances as the "positive" class (Liu et al., 2016). In such a setting, we are concerned with selecting models that work well for precision in the top $k\%$ of

predicted scores (Boyd et al., 2012) rather than optimizing accuracy or area under the ROC curve (AUC-ROC) as often done in "standard" classification problems, which would be suboptimal.

2.1.3. Heterogeneous data sources with strong spatiotemporal patterns

Developing a feature set that adequately represents individuals in policy applications typically entails combining several heterogeneous data sources, often introducing complex correlation structures to the feature space not usually encountered in ML problems used in research settings. For instance, in Bauman et al. (2018), the ML model combines data sources such as criminal justice data (jail bookings), emergency medical services data (ambulance dispatches), and mental health data (electronic case files) to gain a meaningful picture of an individual's state. Additionally, temporal patterns in the data are often particularly instructive, requiring further expansion of the feature space to capture the variability of features across time (number of jail bookings in the last 6 months, 12 months, and 5 years). The combination of such features across a range of domains, geographies, and time frames yields a large (and densely populated) feature space compared to typical structural data-based ML problems we encounter in research settings.

2.2. Socio-technical systems

Typical ML-supported public policy decision-making systems have at least four types of users that interact with ML models at different stages of the process:

1. *ML practitioners* who build the ML components of the system.
2. *High-level decision-makers/regulators* who determine whether to adopt the ML models in their decision-making processes or are responsible for auditing the ML models to ensure intended policy outcomes.
3. *Action-takers* (e.g., social workers, health workers, and employment counselors) who act and intervene based on the recommendation of the ML model. Most policy applications of ML do not involve fully automated decision-making, but rather a combined system of ML model and action-taker that we consider as one decision-making entity. Action-takers often make two types of decisions: (a) deciding whether to accept/override the model prediction for a given entity *(whether to intervene)* and (b) deciding which intervention to select in each case *(how to intervene)*.
4. *Affected individuals* who are impacted by the decisions made by the combined human–ML system.

3. The Role of Explainable ML in Public Policy Applications

Based on our extensive experience working on over 100 such projects in collaboration with governments and nonprofits and through extensive discussions with stakeholders in public policy settings including policymakers, directors of agencies, policy analysts, end users such as counselors and social workers, as well as the public that is impacted, we identify five primary use cases for explainable ML in a public policy decision-making process (see Table 1). For each use case, we identify the end user(s) of the explanations, the goal of the explanations, and the desired characteristics of the explanations to reach that goal for that user. To better illustrate the use cases, we will make use of concrete applications drawn from our work—preventing adverse interactions between police and the public (citation omitted due to blind review)—to serve as a running example. Many applied ML contexts share a similar structure, such as: supporting child welfare screening decisions (Chouldechova et al., 2018), allocating mental health interventions to reduce recidivism (Bauman et al., 2018; Rodolfa et al., 2020), intervening in hospital environments to reduce future complications or readmission (Ramachandran et al., 2020), and recommending training programs to reduce risk of long-term unemployment (Zejnilović et al., 2020).

Illustrative example: Adverse incidents between the public and police officers, such as unjustified use of force or misconduct, can result in deadly harm to citizens, decaying trust in police, and less safety in affected communities. To proactively identify officers at risk for involvement in adverse incidents and

Table 1. *Use cases of explainable ML in public policy applications*

Use case	End users	Intended use of explanations
Model debugging	ML system developers	Uncover and fix errors in the ML pipeline such as leakage or biases
Trust and adoption	Policymakers, regulators, and action-takers	Help users understand how the model makes decisions, evaluate its reasonableness, and sufficiently trust the model for adoption
Whether to intervene	Action-takers	Improve the decision-making system performance by helping action-takers identify correct and unreliable predictions by explaining how the model arrived at individual risk scores
How to intervene	Action-takers	Improve intervention choices by helping action-takers understand factors that contribute to risk
Recourse	Affected individuals	Help affected individuals take action to improve their outcomes in the future or appeal decisions based on inaccurate data

prioritize preventative interventions (e.g., counseling, training, and adjustments to duties), many police departments make use of early intervention systems (EIS), including several ML-based systems (see, e.g., Carton et al., 2016). The prediction task of the EIS is to identify k currently active officers who are most likely to be involved in an adverse incident in a given period in the future (in the next 12 months), where the intervention capacity of the police department determines k. The EIS uses a combination of data sources such as officer dispatch events; citizen reports of crimes; citations, traffic stops, and arrests; and employee records to represent individual officers and generates labels using their history of adverse incidents (Carton et al., 2016).

3.1. Use Case 1: Model debugging

ML model-building workflows are inherently iterative, and one critical piece of this workflow is the continuous feedback provided by sanity checks on the model(s) to see if they *make sense* and are free from errors. A primary goal of explanations at this early stage is to help the system developers identify and correct errors in the models. Common errors such as data leakage (the model having access to information at training/building time that it would not have at test/deployment/prediction time; Kaufman et al., 2011), and spurious correlations/biases (that exist in training data but do not reflect the deployment context of the model) are often found by observing model explanations and finding predictors that should not show up as highly predictive (Caruana et al., 2015; Ribeiro et al., 2016). ML models trained to predict/detect real-world events typically learn from messy data that capture only a partial view of individuals or entities and are highly susceptible to surfacing spurious correlations. Therefore, having additional insight into what the ML model is learning and how it makes decisions through explainable ML can support the model evaluation process. For instance, in Caruana et al. (2015), the authors elaborate on how explanations helped surface such errors in a model trained to identify pneumonia patients with high mortality risk. The model explanations showed that the model assigned low-risk scores to asthma patients because the model did not have access to the information that asthma patients routinely received a more intensive care regimen.

　　Example. In the EIS, an adverse incident gets determined to be *unjustified* a long time after the incident date. When training an ML model with the entire incident record, accidentally using the future *determination state* of the incident can introduce data leakage. In this case, explanations could uncover that a feature such as the *case disposition code* is considered important by the model when it takes a value related

to the determination state and can point the ML practitioner or a domain expert to recognize that information has leaked from the future.

3.2. Use Case 2: Building trust for model adoption

Decision-makers have to sufficiently trust the ML model to adopt and use them in their processes. Trust, in general, is a common motivating theme cited by explainable ML work (Ribeiro et al., 2016; Lipton, 2018; Lundberg et al., 2018a). Furthermore, there has been significant emphasis laid on developing "trustworthy ML" systems of which explainable ML is considered to be a key element (Li et al., 2022). In our experience, trust in human–ML collaboration takes two forms in policy contexts: (a) high-level decision-maker's trust in the model that leads to its adoption and (b) action-taker's trust in the model's predictions that leads to individual actions/interventions. This use case focuses on the former, where the goal of explanations is to help users (policymakers, organizational leadership, etc.) understand and adequately trust the model's overall decision-making process.[2]

The role of the explanation, in this use case, is to help the users understand what factors are affecting the model predictions, as well as the characteristics of individuals that are being scored as high or low risk. Since the user in this instance is not an ML expert but has expertise in the application domain, communicating the explanation in a way that increases the chances of building trust is critical.

Example. In the EIS, the explanations should inform the ranking officer at the PD—who acts as the regulator—of the factors that lead to increasing/decreasing a police officer's risk score (Carton et al., 2016). In that instance, *"A high number of investigations in the last 15 years"* is an interpretable indicator while *"positive first principal component of the arrest code"* is not.

3.3. Use Case 3: Deciding whether to intervene

No ML model makes perfect predictions, especially when predicting rare real-world events. For example, consider an ML model that predicts children and homes at risk of future lead hazards for allocating limited inspection and remediation resources. If only 5% of households have lead hazards, a model that identifies these hazards with a 30% success rate (precision) would provide a significant improvement over a strategy of performing random inspections, but would still be incorrect 70% of the time. In the ideal case, the action-taker in the loop (the lead inspection team) would use their expertise to determine when to follow and act on the model's recommendation and when to override it, resulting in an improved list of k entities. This use case is closely related to the notion of trust we discussed in the above use case, but at the level of individual predictions and with the end user being the action-taker.

Effective explanations can help users, combined with their domain expertise, determine when the model is wrong and improve the overall decisions made by the combined human–ML system. Therefore, the goal of explanations in this use case is to help the action-taker decide *whether to intervene* given the model prediction and its explanations such that the performance of the decision-making system improves (e.g., precision@k in the example above). As the end users are domain experts, the user-interpretability requirement from the above use case holds for the explanations. This use case has been the most commonly studied in explainable ML literature, albeit in non-policy settings. For instance, in Ribeiro et al. (2016), the authors study whether explanations generated from their method (LIME) can highlight the predictors that contributed to the prediction and assist users to identify "unreasonable" predictions through a simulated-user study, Lundberg et al. (2018b) studied the ability of explanations to assist physicians to detect hypoxemia risk during surgery, Jesus et al. (2021) studied the ability to assist fraud analysts to detect credit card fraud with explanations of ML predictions.

---

[2] It is important to note that explainability is not the only factor that affects user trust. In a policy context, factors such as (a) stability of predictions, (b) the training users have received, and (c) user involvement in the modeling process, also impacts user trust (Ackermann et al., 2018).

Example. In the EIS, if an explanation exists for each officer in the top-*k*, outlining the factors contributing to the risk score, the internal affairs division—who decides *whether to intervene*—can use those explanations to determine the reliability of the model's recommendation to act on it or override it.

3.4. Use Case 4: Deciding how to intervene

While ML models can help identify entities that need intervention, they often provide little to no guidance on selecting interventions. For instance, consider a model that predicts students' risk of not graduating high school on time. A student might be at risk due to several reasons, such as struggling with a specific course, bullying, transportation issues, health issues, or family obligations. Each of those reasons would require a different type of assistive intervention. ML explanations can highlight the predictors that contribute to the risk score and could help a teacher or other domain expert identify the reasons behind the predicted high risk of a student.

Therefore, in this use case, the goal of the explanation is to help the action-taker decide *how to intervene* and often choose among one of many possible interventions available. While typical ML explanations are not truly causal, the factors highlighted in the explanation can provide valuable information to a domain expert in choosing interventions. While there have not been studies on this use case for policy administration decisions to the best of our knowledge, there have been a few efforts in other domains where researchers have investigated using explainable ML for supporting recommending actions. For instance, Afzaal et al. (2021) showed that explanations of student performance predictions can be used to recommend actions to students in self-regulated learning, Albreiki (2022) studied how explanations from ML can be used to recommend remedial actions to low-performing students with the goal of improving learning outcomes, and Sajja et al. (2021) demonstrated the use of explainable ML predictions of consumer behavior in helping fashion designers plan for new products.

Example. Consider an officer flagged by the EIS, with an explanation indicating that the model is prioritizing features related to the type of dispatches the officer was assigned to in the last few months. Upon further inspection of the data, it can be seen that the officer had been dispatched to high-stress situations regularly. In this instance, a possible intervention is reassigning duties or putting them on low-stress dispatches after a series of high-stress dispatches.

3.5. Use Case 5: Recourse

When individuals are negatively impacted by ML-aided decisions, providing them with a concrete set of actionable changes that would lead to a different decision is critical. This ability of an individual to affect model outcomes through actionable changes is called recourse (Ustun and Rudin, 2019). While recourse has been studied independently from explainable ML (Ustun and Rudin, 2019; König et al., 2021), ML explanations have the potential to help individuals seek recourse in public policy applications (Wachter et al., 2018; Karimi et al., 2020, 2021b).

In this use case, there are two explanation goals: (a) helping the user understand the reasons behind the current decision, allowing them to discover any inaccuracies in the model and/or data and dispute the decision, and (b) helping the user identify the set of actionable changes that would lead to an improved decision in the future. As the user in this use case is the affected individual, the explanations that indicate the reasons behind the decisions should be mapped to a domain that is understandable by the individual. Furthermore, the explanations should recommend feasible and actionable changes (e.g., reducing age by 10 years vs. reducing debt).

Example. In the EIS, the affected individual is the flagged officer. If the officer is provided with explanations indicating the reasons behind the elevated risk score and actionable changes that could reduce their risk score, they could either point to any inaccuracies or take measures themselves (in addition to the intervention by the PD) to reduce the risk score.

4. Current State of Explainable ML

In this section, we summarize the existing approaches in explainable ML. It is worth noting that the intention here is not to provide an in-depth and comprehensive literature review but rather a broad view of existing approaches and discuss how they apply to the public policy settings described above. We refer readers to Adadi and Berrada (2018), Guidotti et al. (2018), Arya et al. (2019), Molnar (2019), and Bhatt et al. (2020b) for more comprehensive reviews of existing work.

4.1. Existing work in explainable/interpretable ML

Existing approaches broadly fall into two categories: (a) inherently interpretable ML models and (b) post hoc methods for explaining opaque ML models.[3] ML explanations take two forms: (a) explaining individual predictions (local explanation) and (b) explaining the overall behavior of the models (global explanation). Typically, local explanations are intended to help users understand *why* the model arrived at the given prediction for a given instance, while global explanations explain *how* the model generally behaves (Plumb et al., 2018). Table 2 summarizes the existing approaches.

4.1.1. Inherently interpretable ML models

Inherently interpretable ML models are designed such that an end user could understand its decision-making process (Lakkaraju et al., 2016; Rudin, 2019). In a policy context, an interpretable model could allow a user to (a) understand how the model calculates a risk score (global explanation) and (b) understand what factors contributed to the predicted risk score (local explanation) for a given instance. Several efforts have focused on developing interpretable models for policy domains, such as those for healthcare and criminal justice (Caruana et al., 2015; Zeng et al., 2017). These include sparse linear models (Ustun et al., 2013; Ustun and Rudin, 2019), sparse decision trees (Hu et al., 2019), generalized additive models (Hastie and Tibshirani, 1990; Lou et al., 2012, 2013; Caruana et al., 2015), and interpretable decision sets (Lakkaraju et al., 2016).

Interpretable models often rely on carefully curated representations of data with meaningful input features (Rudin, 2019), often through discretization or binary encoding (Caruana et al., 2015; Lakkaraju et al., 2016; Ustun and Rudin, 2019). Distilling complex data spaces into a handful of optimally discretized and meaningful features can entail extensive effort and optimization of its own. While careful feature preparation is indispensable in any ML application, regardless of the employed ML algorithm complexity, distilling complex and heterogeneous feature spaces typically found in policy settings into a handful of simple features can prove to be particularly challenging.

4.1.2. Post hoc methods for explaining black-box ML models

Post hoc methods derive explanations from already trained black-box/opaque ML models. As post hoc methods do not interfere with the model's training process, they enable the use of complex ML models to achieve explainability without the risk of sacrificing performance. However, as black-box ML models are often too complex to be explained entirely, post hoc methods typically derive an approximate explanation (Gilpin et al., 2018; Rudin, 2019), which makes ensuring the fidelity of the explanations to the model a key challenge in this work. Unlike inherently interpretable models, local and global explanations for opaque complex ML models require different methods. For both types of explanations, both model-specific and model-agnostic methods exist in the literature.

Post hoc local explanations. A local explanation in a typical public policy problem is used to understand which factors affected the predicted risk score for an individual entity. The most common format of local explanation is feature attribution—also known as feature importance or saliency—where each input feature is assigned an importance score that quantifies its contribution to the model prediction

---

[3] Note that model opacity can be a reflection of either (a) the model being too complex to be comprehensible, or (b) the model being proprietary (Rudin, 2019). In this paper, we focus on opacity created through model complexity.

Table 2. *A summary of existing approaches for explainable ML*

| | Post hoc methods | | | | |
| | Local | | Global | | |
Approach	Model-agnostic	Model-specific	Model-agnostic	Model-specific	Interpretable models
Sparse models[a]					✓
Decision rules and sets[b]					✓
Gen. additive models[c]					✓
Local surrogate models[d]	✓	✓			
Permutation (Shapley values)[e]	✓	✓		✓	
Global rule extraction[f]		✓	✓		
Gradient-propagation[g]		✓			
Counterfactuals[h]	✓	✓			
Example based[i]	✓		✓		

[a]Ustun et al. (2013, 2019), Zeng et al. (2017), Hu et al. (2019).
[b]Letham et al. (2015), Lakkaraju et al. (2016), Ye et al. (2019).
[c]Hastie and Tibshirani, (1990), Lou et al. (2012), Caruana et al. (2015).
[d]Ribeiro et al. (2016), Plumb et al. (2018).
[e]Lundberg and Lee, (2017), Lundberg et al. (2018a, 2018b, 2020).
[f]Tsukimoto, (2000), Ribeiro et al., (2018).
[g]Baehrens et al. (2010), Simonyan et al. (2013), Zeiler and Fergus, (2014), Bach et al. (2015).
[h]Wachter et al. (2018), Ustun et al. (2019), Mothilal et al. (2020), Poyiadzi et al. (2020).
[i]Kim et al. (2016), Koh and Liang, (2017).

(Baehrens et al., 2010; Bhatt et al., 2020b). Several approaches exist for deriving feature importance scores such as: fitting an interpretable surrogate model (linear classifier) around a local neighborhood of the instance in question (Ribeiro et al., 2016; Plumb et al., 2018); feature perturbation- based methods for approximating each feature's importance using game-theoretic Shapely values (Lundberg and Lee, 2017; Lundberg et al., 2018a); and gradient-based techniques (Simonyan et al., 2013; Zeiler and Fergus, 2014; Bach et al., 2015). Among these approaches, methods such as LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), and SA (Zeiler and Fergus, 2014) are model-agnostic methods, whereas LRP (Bach et al., 2015), deconvolution (Simonyan et al., 2013), and TreeSHAP (Lundberg et al., 2018b) are model specific methods. MAPLE (Plumb et al., 2018) stands out among these methods as it can act both as an inherently interpretable model as well as a model-specific post hoc local explainer.

Other approaches such as influence functions (Koh and Liang, 2017) as well as prototypes and criticisms (Kim et al., 2016) make use of data instances, rather than features, to provide local explanations. A special form of example-based explanation is counterfactual explanations, which seeks to answer the following question: *"what's the smallest change in data that would result in a different model outcome?"* (van der Waa et al., 2018; Wachter et al., 2018; Molnar, 2019; Barocas et al., 2020; Mothilal et al., 2020; Karimi et al., 2021b). In a top-k setting, the *change in outcome* can be the inclusion versus exclusion of the individual from the top-k list. Counterfactual explanations can provide insight into *how to act* to change the risk score, supplementing the feature attribution methods that explain *why* the model arrived at the risk score.

Post hoc global explanations. A global explanation in a typical policy problem would be a summary of factors/patterns that are generally associated with high-risk scores, often expressed as a set of rules (Plumb et al., 2018; Ribeiro et al., 2018). Global explanations should enable the users to accurately predict, sufficiently frequently, how the model would behave in a given instance. However, deriving global explanations of models that learn highly complex nonlinear decision boundaries is very difficult (Ribeiro et al., 2016). As a result, the area of deriving post hoc global explanations is not as developed as local explanation methods.

Some approaches for global explanations from black-box ML models include (a) aggregation of local explanations (Lundberg et al., 2020), (b) global surrogate models (Frosst and Hinton, 2017), and (c) rule extraction from trained models (Tsukimoto, 2000). A noteworthy contribution to deriving globally faithful explanations is ANCHORS (Ribeiro et al., 2018). ANCHORS identifies feature behavior patterns that have high precision and coverage in terms of their contribution to the model predictions of a particular class. Methods proposed by Lundberg et al. (2020) and Ribeiro et al. (2018) are model-agnostic and methods presented by Frosst and Hinton (2017) and Tsukimoto (2000) are model-specific.

4.2. Capabilities of existing explainable ML methods and public policy use cases

In this section, we characterize the established capabilities of the existing explainable ML methods classes with respect to the use cases we identified. To rank capabilities, we use a three-point scale that is based on the level of evidence that existing method evaluations demonstrate for an individual use case (see Table 3). Highlighting the multi-faceted nature of evaluating explainable ML methods, Doshi-Velez and Kim (2017) called for more rigorous approaches in the field and mapped these evaluation studies into a three-tiered framework: (a) functional-grounded evaluation, where the intrinsic qualities of the explanation are evaluated purely through algorithmic means, for example, the fidelity of the explanation to the underlying ML model is a commonly used metric in functional-grounded evaluation, (b) human-grounded evaluation, where the utility of the explanations are assessed using proxy users or simplified tasks, for example, users from AMT performing a task such as simulating the ML model's prediction given the data and explanation is a commonly used human-grounded evaluation setup, and (c) application-grounded evaluation, where the utility of explainable ML is tested at helping real users (domain experts) perform a real-world task. The three-point scale that we present is primarily based on human-grounded and application-grounded evaluations as we are interested in highlighting the proven utility of explainable ML

Table 3. Capabilities of existing methods with respect to the public policy use cases

Use case	Post hoc local	Post hoc global	Interpretable models
Model debugging[a]	★★☆	★★☆	★★☆
Trust and adoption[b]	★☆☆	★☆☆	★☆☆
Whether to intervene[c]	★★☆	×	★☆☆
How to intervene[d]	★☆☆	×	★☆☆
Recourse[e]	★★☆	×	×

Note. Please note that the references cited in the table indicate the publications the highest rating received is based on.
[a]Caruana et al. (2015), Ribeiro et al. (2016).
[b]Ribeiro et al. (2016), Ribeiro et al. (2018), Buçinca et al. (2020), Hase and Bansal, (2020), Jacovi et al. (2021).
[c]Ribeiro et al. (2016), Lundberg and Lee, (2017), Lundberg et al. (2018a), Jesus et al. (2021).
[d]Lou et al. (2012), Bach et al. (2015), Lakkaraju et al. (2016), Ribeiro et al. (2016), Lundberg and Lee, (2017), Plumb et al. (2018), Lundberg et al. (2018a), Hu et al. (2019), Ustun et al. (2019).
[e]Ustun et al. (2019), Karimi et al. (2020, 2021b, 2021a), Mothilal et al. (2020).

methods in the identified use cases. Our goal for this ranking is to highlight where the established capabilities in the field fall short of the needs of the use cases based on our research and our experience implementing and evaluating them. We use the three broad method categories—post hoc local methods, post hoc global methods, and inherently interpretable models—for this ranking and assign a rating to the whole group if at least one method in the group satisfies the requirements. We define the three-point scale as follows:

★☆☆: Methods are potentially applicable to the use case. However, we have not found any human-grounded or application-grounded studies where any method in the class is directly evaluated on the use case and shown to be effective.
★★☆: Some evidence of efficacy in the use case exists through evaluations on simplified/proxy problems and proxy users (human-grounded evaluations). However, no application-grounded studies exist where the efficacy of any method is empirically validated through a well-designed user study in a real-world setting where real users are performing a real task.
★★★: At least one method in the group is validated with an application-grounded evaluation on the use case with a well-designed user study, which implements the method on a real task, uses real data, presents explanations to real users of the system, and empirically demonstrates the method's efficacy at improving outcomes of interest.
×: Methods in the group are not applicable to the use case.

The discussion below summarizes how existing work maps to each use case and our assessment of the status of current work with respect to these applications. It is worth noting that inherently interpretable models are potentially applicable to all the use cases. Therefore, we focus on the post hoc methods in the summaries below.

4.2.1. Model debugging

Methods for both local and global post hoc explanations are potentially useful in this use case. Global explanations could help identify errors in overall decision-making patterns (e.g., globally important features can help identify data leakage), and local explanations can help to uncover errors in individual predictions.

Although some recent work lends evidence for the utility of explanations in discovering model errors (Caruana et al., 2015; Ribeiro et al., 2016; Adebayo et al., 2020; Abid et al., 2022), the efficacy of these methods is not empirically validated through well-defined user trials in real-world applications. For instance, Ribeiro et al. (2016) demonstrate how LIME explanations could help users identify model errors through a simplified image classification task and a text classification task. While these studies show that users performed better with the availability of explanations, we argue the fact that both classification tasks

were simplified by introducing errors to the model, and explanations were presented to users from AMT oversimplifies the task and deviates the experimental setting from real-world applications, rendering the efficacy of the method to be inconclusive.

4.2.2. Model trust and adoption

As with model debugging, both global and local explanation methods are potentially applicable. However, as the end user is the domain expert, explanations will need to be extended beyond feature attribution while preserving fidelity to what the model has learned. While existing methods discuss user trust as a broad goal, to the best of our knowledge, their ability to help regulators or decision-makers adequately trust ML models is not demonstrated through well-defined evaluations or user trials. The experimental work on the notion of trust has relied on subjective, self-reported measures of trust in performing a simplified task (Ribeiro et al., 2016; Weitz et al., 2019; Buçinca et al., 2020). However, Jacovi et al. (2021) in their effort of formalizing the notion of trust in ML, argue that simply asking the user whether they trust the model for a simple task does not evaluate the notion of trust in AI, as the users are not assuming any risk, and they argue that relying on an AI with assumed risk is a prerequisite for trust. A proxy task that is often employed in the evaluation of explainable ML is "forward-simulation" (Ribeiro et al., 2016, 2018; Doshi-Velez and Kim, 2017), that is, a person predicting the ML model's outcome given the input and explanation. This ability to accurately anticipate the model's output is considered a proxy signal of trust (Hase and Bansal, 2020; Jacovi et al., 2021). However, despite these initial efforts, to the best of our knowledge, there have not been experimental efforts that study how existing explainable ML methods impact the notion of trust related to model adoption in decision-making processes, and how it relates to the societal outcomes of interest.

4.2.3. Improving decision-making system performance

Feature attribution-based local explanations are potentially applicable to provide the necessary information to the user. However, feature attribution alone may not be sufficient. Users may need more contextual information such as *How does the instance fit into the training data distribution? How does the model behave for similar examples? And what factors did it rely on for those predictions?* To that end, there have been some efforts to present visual summaries of explanations to the user (Lundberg and Lee, 2017; Ribeiro et al., 2018; Lundberg et al., 2018b) which could potentially be useful in this use case. Therefore, available local explanation methods do provide a good starting point.

In contrast to the above use cases, there have been a couple of instances where explainable ML methods were tested using an experimental setting consisting of a real task, real data, and real users. Lundberg et al. (2018b) studied the ability of an explainable ML system to assist anesthesiologists to detect hypoxemia risk during surgery for proactive intervention. They showed that their system—Prescience—armed with an ML model and SHAP explanations was able to outperform the anesthesiologists in identifying real-time hypoxemia risk. However, their experiment failed to isolate the marginal effect of the explanations by failing to compare the performance of the ML model + explanations to just the ML model. Therefore, while the combined system with predictions and explanations outperformed the domain expert, it was not possible to isolate whether the effects were due to the ML model prediction alone or due to the combined system. Jesus et al. (2021) studied the impact of presenting ML explanations from three local post hoc explainable ML methods to fraud analysts for assisting fraud detection in credit card transactions. While they organize the experiment to isolate the incremental impact of explanations by running appropriate experiment arms, they make several simplifying assumptions in their experimental design. For instance, they resample the data in their experiment to reflect a 50% rate, whereas, in the real-world context, fraud is a significantly rarer event. Furthermore, they measure the effectiveness of decisions using confusion matrix-based metrics, assuming that all transactions are of the same value, an assumption that is not valid in a real-world business setting. Therefore, while Jesus et al. (2021) have taken steps in the right direction, we argue that their experimental setup still does not reflect the deployment context adequately.

Despite the existence of user studies conducted with real data and real users, the significant limitations of these studies to date mean they fail to provide conclusive evidence that existing explanation methods are effective in assisting humans to make improved decisions in this use case.

4.2.4. How to intervene

As the intervention determinations are often individualized, local explanation methods are potentially applicable for generating the reasons behind the risk score. As with the above use case, users may need more contextual information to supplement the local explanations such as: *how the instance fits into the training data distribution*, and *intervention history* for *similar—w.r.t.* data and *w.r.t.* explanation—individuals. To the best of our knowledge, there is a lack of evidence in the existing body of work on the efficacy of using these local explanation methods for informing intervention selection.

4.2.5. Recourse

Feature attribution-based local explanations are potentially applicable for deriving *reasons* behind the decision, and counterfactual explanations can be useful in explaining how to improve the outcomes. The focus of algorithmic recourse work has been on using counterfactual explanations. As simple counterfactual explanations do not guarantee explanations with actionable changes, there has been a range of approaches proposed for deriving counterfactual explanations that are *diverse*, *sparse*, *plausible*, and *actionable* (Ustun et al., 2019; Karimi et al., 2020, 2021b; Mothilal et al., 2020; Poyiadzi et al., 2020; Upadhyay et al., 2021). Karimi and colleagues provide a survey of methods for algorithmic recourse including in Karimi et al. (2021a). Evaluating methodologies effectively has been a challenge for algorithmic recourse methods and Karimi and colleagues call for better benchmarks (Karimi et al., 2021a). Therefore, existing method evaluation largely relies on theoretical guarantees and demonstrations using popular experimental datasets such as adult-income, German credit lending, and COMPAS. While those datasets have some connection to the real world, they are not reflective of datasets that most real-world organizations in policy settings have, especially in terms of richness, complexity, and spatiotemporal patterns. Studies that empirically validate the efficacy of these methods are still lacking. Since there are no user studies for any class of models, we rate post hoc local methods with one star.

5. Gaps and Proposed Research Directions

In this section, we use the mapping between methods and use cases to identify gaps in existing explainable ML research compared to the needs of real-world public policy problems and propose a research agenda to fill those gaps. We believe that bridging these research gaps is critical for applying ML to social problems and if we are to safely deploy ML systems that lead to effective policy decisions and have a positive and lasting impact on society.

5.1. Gap 1: Capabilities of existing methods not adequately evaluated in real-world contexts

The most pronounced gap between explainable ML research and the policy use cases is the lack of evidence of method efficacy established through rigorous application-grounded evaluation studies (in our review, we failed to find any study that met the criteria to achieve a three-star rating in Table 3). The most common approach to evaluating explainable ML has been to assess the quality of the explanation (the artifact produced by the method) through functional-grounded evaluations. Almost all user studies take a human-grounded approach where nonexpert users (e.g., users from AMT or users in research settings) perform simplified/proxy tasks such as "forward simulation."

A growing body of work that empirically demonstrates the limitations of the functional-grounded and human-grounded approaches has begun to appear recently. Hase and Bansal (2020) describe an experiment where users were asked to perform the task of "forward simulation" and subjectively evaluate the

quality of explanations (measuring a concept of "simulatability"). Although simulatability seems unlikely to reflect real-world use of explanations, it is notable that the authors found essentially no relationship between the human-grounded subjective assessments of explanation quality and how users performed on this task. Similarly, Buçinca et al. (2020) compared three proposed measures of explainable ML: subjective user assessments, user performance on a proxy task (predicting model scores based on explanations, similar to the study by Hase and Bansal), and performance on a decision-making task (assessing the nutritional content of different plates of food). Their results indicated that performance on neither the subjective assessments nor the proxy task generalized to performance on the actual decision-making task, highlighting the risks of relying too heavily on these simplified evaluations. We argue that functional-grounded and human-grounded evaluations are not sufficient to establish the utility of explainable ML methods in domains such as public policy where ML systems learn from highly complex, heterogeneous, and messy data spaces and assist consequential decisions. In this work, we are interested in evaluating explainable ML methods on their ability to improve a societal outcome of interest. Functional-grounded metrics such as fidelity do not guarantee the usefulness of explanations and designing proxy tasks for human-grounded evaluations that capture the nuances and complexities of public policy problems can be challenging.

5.1.1. Guidelines for adequate evaluation of explainable ML in policy contexts

Given the limitations of functional-grounded and human-grounded approaches, evaluation studies of ML explanations in policy contexts should focus on application-grounded approaches. Doshi-Velez and Kim (2017) define an application-grounded evaluation as a study where domain experts perform the intended task. We further extend those requirements and argue that an adequate application-grounded evaluation of an explainable ML method in a policy context cannot exist in the absence of four key elements: (a) A real policy task and related metrics, (b) users who perform the task in the real world, (c) real-world data related to the task that captures the complexities and nuances of the task, and (d) a robust inference strategy that allows conclusions on the incremental impact of explanations. Unfortunately, the relatively small number of application-grounded explainable ML studies that incorporated some aspects of practical evaluation have consistently lacked at least one (and in many cases multiple) of the elements necessary to offer conclusive evidence of real-world efficacy. We elaborate on each of these elements and how existing work violates those requirements below:

Defining the task. The task here entails the decision a user would make, the goals of the decision-making process, and the metrics that evaluate its success. It is imperative to pick tasks that align with a well-defined policy/operational goal and metrics that directly measure the success of the task, going beyond general-purpose metrics such as ROC-AUC, accuracy, and F1-score. For instance, consider the study conducted by Jesus et al. (2021) in a credit-card fraud detection context. While the authors conduct the study with professional fraud analysts performing fraud detection on real-world credit card transactions, they make a simplifying assumption on selecting metrics and choose decision accuracy and other confusion matrix-based metrics. In the context of e-commerce transactions, we argue that the goal is to maximize revenue/profit, and the metric should factor in the transaction value and the relative costs of the two types of errors. Choosing accuracy as the metric ignores both these nuances and thus violates the first requirement of a study. In Poursabzi-Sangdeh et al. (2021), the authors used real-world data and a large cohort of real users. However, they choose to use the task of real estate valuation, and how it relates to a decision and an outcome of interest and the utility of explanations in achieving that goal was not established.

Recruiting users. Although their time is often scarce, the domain expert users who act on model outputs must be involved in the evaluation process to ensure it reflects the actual deployment scenario. As the interaction between model predictions, explanations, and users' domain expertise will dictate the performance of the system, substituting inexperienced users (for instance, from AMT) provides little insight into how well explanations will perform in practice and the results generated (Lou et al., 2013;

Lundberg and Lee, 2017; Lundberg et al., 2018a; Hu et al., 2019) may not correlate with results in actual deployment with real users. We argue that even if a study implements explainable ML methods on real-world problems and data that capture the nuances and complexities of the domain, the absence of evaluation with domain experts leads to uncertainty of method efficacy. For instance, Caruana et al. (2015) and Zeng et al. (2017) describe evaluations of their inherently interpretable models on a real problem with a clear goal and real-world data. However, both studies failed to conduct studies with actual users of the system to evaluate the incremental utility of explanations in terms of performance improvement in the task.

Data. To capture the nuances and characteristics of applying ML to a policy area in practice, the use of data from the problem domain is essential. This is particularly important with evaluating explainable ML methods, as simplified or synthetic datasets might provide an overly optimistic evaluation of their ability to extract meaningful information. Unfortunately, most of the work in this area fails to meet these criteria, focusing only on benchmark ML problems and datasets (e.g., image classification on MNIST data, newsgroups data; Bach et al., 2015; Ribeiro et al., 2016, 2018; Shrikumar et al., 2017). While benchmark problems play a crucial role in explainable ML method development and refinement, these problems are far removed from deployment contexts encountered in public policy settings and thus fail to provide convincing evidence of method effectiveness in informing the decisions of domain experts. It is important to note that even in studies that do use data from the problem domain, seemingly trivial simplifying assumptions can violate this requirement. For instance, in Jesus et al. (2021), the authors adjust the class distribution to artificially create a 50–50 distribution to remove the "class imbalance" problem, whereas the actual fraud rate is around 15%. This simplification can impact the expected prior beliefs of fraud analysts and the findings of the study.

Defining the inference strategy. In addition to setting up problems, data, and users consistent with the deployment context, one aspect where existing studies have faltered is the design of the inference strategy of the experiment. It is essential to design the inference strategy to support conclusions on the incremental impact of the explanations in the context. A robust inference strategy entails evaluating the appropriate hypotheses, appropriate control/treatment groups, sufficient sample sizes to preserve statistical power, and analytical methodologies that capture the uncertainties in data. Consider the study conducted by Lundberg et al. (2018b), where the authors evaluate an explainable ML system on its ability to support anesthesiologists detect hypoxemia during surgeries. The authors implemented the method on the actual task, recruited domain expert users, and used historical data captured during surgeries. Unfortunately, the study fails to evaluate the incremental impact of the explainable ML method as it does not evaluate the appropriate hypotheses. The authors compare the performance between users making decisions only using data and users making decisions with the help of data, ML prediction, and explanation. The study concludes that users perform better when they have access to ML predictions and explanations compared to only having access to the data. However, the authors fail to compare the performance of users with access to data and ML prediction (no explanations) to the user performance with all three pieces of information. We argue that this is an essential component of an experimental design aimed at evaluating the incremental impact of explainable ML, and this oversight prevents us from attributing that performance increment to the explanations.

As we can see, even the relatively few application-grounded explainable ML studies that have incorporated some aspects of practical evaluation have consistently lacked at least one (and in some cases multiple) of the elements necessary to offer conclusive evidence of real-world efficacy (see Table 4). This gap is particularly acute in the context of public policy problems, given the characteristics (see Section 2) that set them apart from ML problems we encounter in research settings. Therefore, we argue that for implementing explainable ML systems in policy settings, there should be concerted efforts from ML and policy practitioners to conduct well-designed application-grounded evaluation studies.

Table 4. *Analyzing the existing handful of application-grounded evaluation studies with respect to the proposed desiderata*

Study	Real task		Real data	Real users	Inference
	Decision	Metrics			
Jesus et al. (2021)	✓	✗	✗	✓	✓
Lundberg et al. (2018b)	✓	✓	✓	✓	✗
Poursabzi-Sangdeh et al. (2021)	✗	✗	✓	✓	✓
Caruana et al. (2015)	✓	✓	✓	✗	✗
Zeng et al. (2017)	✓	✓	✓	✗	✗

Note. Please note that we include studies that at least satisfies one requirement.

5.1.2. An example evaluation study design that satisfies the desiderata

To make the desiderata more concrete, this section provides an overview of an experimental design for evaluating an explainable ML method in a policy context that implements the four necessary elements we presented. While this example describes a setting that uses a post hoc explainable ML method, the same setup could be used to evaluate an inherently explainable model. The policy problem presented here is similar to the one discussed by Bauman et al. (2018).

The policy problem. The Mental Health Center (MHC) of a mid-sized suburban county is establishing a program to conduct proactive mental health outreach to individuals at risk of criminal justice involvement due to unmet behavioral health needs. The MHC has the resources to perform outreach and assist about 100 individuals each calendar month. The broad policy goal is to minimize repeated criminal justice involvement among county residents, and the specific goal of the human–ML collaborative system would be to maximize the efficient use of county resources.[4]

ML model. In order to prioritize individuals for outreach, a ML-based predictive model is used. Among the county residents who were released from jail in the last 2 years, the model predicts the risk of each individual being booked into jail within the next 12 months and maps them into an ordinal scale.

Users: Mental health clinicians of the Mobile Crisis Response Team (MCRT) will act on the predictions of the ML model by conducting outreach and offering appropriate services to these individuals based on their needs.

Task and metrics. Given an individual that is deemed to be at risk of future criminal justice involvement (by the ML model), the MCRT clinician is tasked with verifying the risk and selecting them for mental health outreach. In different experimental conditions, the clinicians would have different information pieces at their disposal to make this decision. Since the goal is maximizing efficiency, the objective is to correctly identify people who are actually at risk. Therefore, the metric we should be optimizing for should capture how accurate the MCRT clinicians are at identifying the individuals for outreach, that is, given that an individual is selected for outreach, maximizing the probability that they are actually at risk of being booked into jail the following year, which can be captured by maximizing the positive predictive value (PPV)/precision.

Data: The ML model learns from historical data from the county jail, emergency medical data, and behavioral health service involvement data at the individual level and makes predictions about future criminal justice involvement. The study uses historical data to train and evaluate the model as well as to evaluate the task performance of the MCRT clinicians.

Explainable ML method. The related use case for the task is Use Case No. 03 (deciding whether to intervene) and we use a feature attribution-based post hoc model agnostic explanation method for the evaluation (e.g., SHAP and LIME).

---

[4] It is worth noting that in a typical project, the efficiency goals are often coupled with an equity goal, but for simplicity, we narrow the scope of this example down to focus on efficient resource allocation.

Inference strategy. In this trial, we are interested in learning whether the explainable ML method is effective in helping MCRT clinicians correctly pick individuals for mental health outreach. At a minimum, the trial should evaluate the following hypotheses:

1. MCRT clinicians select individuals for outreach at a higher precision when presented with ML predictions than when they only have access to data.
2. MCRT clinicians select individuals for outreach at a higher precision when they have access to explanations of ML predictions than when they have access only to ML predictions.

In order to evaluate these hypotheses, we need to create three experimental groups/arms where the clinicians have access to different levels of information:

1. Clinicians have access only to the data of the individuals (Group 1).
2. Clinicians have access to the data, and the predicted risk score, for example, a calibrated probability (Group 2).
3. Clinicians have access to the data, predicted risk score, and explanations generated by the post hoc method for the prediction (Group 3).

In the ideal case, we would need a large enough user base where the unit of randomization could be the clinicians, and we could randomly assign clinicians to the three groups. However, in policy applications, having access to a domain expert user base of significant size to enable sufficient statistical power is rare. In that case, we could design the trial in stages like Jesus et al. (2021) did and randomize at the level of data instances (an individual at a specific point in time). While this limits the hypotheses we can evaluate (e.g., how clinicians with different levels of experience use and interact with ML explanations), we can still measure the efficacy of an explanation method in the use case. With this setup, we can compare Group 2 against Group 1 to evaluate the first hypothesis and similarly compare Group 3 against Group 2 for the second.

5.1.3. *Importance of evaluating the performance–explainability trade-off (if any) for inherently interpretable models*

In addition to evaluating the effectiveness of ML explanations in helping domain experts in public policy settings, it is important to assess the viability of inherently interpretable models in terms of predictive performance. As inherently interpretable models rely on carefully curated input features, exploring the trade-off between performance and scalability in practice is crucial to ascertain their broad applicability. To that end, the models should be implemented on several real policy problems, evaluating: (a) the trade-off between feature preparation efforts and performance and (b) their ability to generalize to future data under strong temporal dependencies.

The prospect of inherently interpretable ML models that are intelligible without sacrificing predictive performance certainly holds considerable appeal. However, their current implementations are limited to a handful of practical contexts. To understand potential trade-offs in practice, we must rigorously test these models against more opaque models across problem domains. Even if there are performance limitations, there may be critical applications where the intelligibility of the model cannot be compromised and some applications where there could be built-in guardrails to protect against unintended harm. Understanding the limitations of methods through experimentation will help practitioners make more informed implementation decisions and support the complete spectrum of use cases.

5.2. *Gap 2: Existing methods are not explicitly designed for specific use cases*

As discussed above, existing methods are developed with loosely defined or generic explainability goals (e.g., transparency) and without well-defined context-specific use cases. As a result, methods are

developed without understanding the specific requirements of a given domain, use case, or user base, resulting in a lack of adoption and suboptimal outcomes.

While several existing methods may be applicable for each use case, their effectiveness in real-world applications is not yet well-established, meaning this potential applicability may fail to result in practical impact. As more methods are rigorously evaluated in practical, applied settings as suggested above, gaps in their ability to meet the needs of these use cases may become evident. For instance, model-agnostic methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) are capable of extracting input feature importance scores for individual predictions from otherwise opaque models. However, it is unclear whether they can address needs such as generating explanations that are well-contextualized and truly interpretable by less technical users without sacrificing fidelity (e.g., to help a domain expert identify unreliable model predictions or an affected individual seek recourse).

6. Conclusion

Despite the development of a wide array of explainable ML methods, their efficacy in improving real-world decision-making systems is yet to be sufficiently explored. In this article, we sought to characterize and understand this gap in the present literature in hopes that this effort can help structure future evaluations of these methods to better address their practical utility. First, we identified the primary set of use cases for ML model explanations in the ML-aided public policy decision-making pipeline: (a) model debugging, (b) regulator trust and model adoption, (c) deciding whether to intervene, (d) deciding how to intervene, and (e) recourse. For each use case, we defined the goals of an ML explanation and the intended end user. Then, we summarized the existing approaches in explainable ML and identified the degree to which this work addresses the needs of the identified use cases. We observed that, while the existing approaches are potentially applicable to the use cases, their utility has not been thoroughly validated for any of the use cases through well-designed empirical user studies.

Two main gaps were evident in the design and evaluation of existing work: (a) methods are not sufficiently evaluated in real-world contexts and (b) they are not designed and developed with target use cases and well-defined explainability goals in mind. In response to these gaps, we proposed several research directions to systematically evaluate the existing methods with problems with real policy goals, real-world data, and domain experts.

A key aim of this article is to connect the ML research community that develops explainable ML methods to the problems and needs of the public policy and social good domains. As computer scientists who develop and apply ML algorithms to social/policy problems in collaboration with government agencies and nonprofits, we are ideally and uniquely positioned to understand both the existing body of work in explainable ML and the explainability needs of the domains such as public health, education, criminal justice, and economic development. Two main factors motivated us to compile this discussion: (a) despite the existence of a large body of methodological work in explainable ML, we failed to identify methods that we could directly apply to the problems we were tackling in the real world and (b) the frequent conversations initiated by our colleagues in the ML research community on how their methods could be better suited and developed for real-world ML problems.

We strongly believe that explainable ML methods will prove to be a critical component of ML systems that are designed for policy and societal problems, where high-stakes decisions with significant impacts on people's lives create a moral imperative for these systems to perform well across all five use cases we discuss. As such, there is considerable potential for explainable ML to have a broad positive impact on society through these applications, but it will only have this impact if we design and develop these methods for explicitly defined use cases and evaluate them in a way that demonstrates their effectiveness on those use cases. Therefore, the goal of this article was not to develop new algorithms, nor was it to conduct a thorough survey of explainable ML work (since there are already several excellent articles on that topic). Rather, our goal was to take the necessary first steps to bridge the gap between methodological work and real-world needs. We hope that this discussion will help the ML research community collaborate with the Policy and HCI communities to ensure that existing and newly proposed explainable ML

methods are well-suited to meet the needs of the end-users to give them the confidence to implement and deploy them in systems that can benefit society.

References

Abid A, **Yuksekgonul M and Zou J** (2022) Meaningfully debugging model mistakes using conceptual counterfactual explanations. In Chaudhuri K, Jegelka S, Song L, Szepesvari C, Niu G and Sabato S (eds), *Proceedings of the 39th International Conference on Machine Learning*, volume *162* of *Proceedings of Machine Learning Research.* Cambridge, MA: PMLR, pp. 66–88, 17–23 July 2022.

Ackermann K, **Naveed H**, **Bennett J**, **Walsh J**, **Rivera AN**, **Defoe M**, **De Unánue A**, **Lee SJ**, **Cody C**, **Haynes L and Ghani R** (2018) Deploying machine learning models for public policy: A framework. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery, pp. 15–22, July 2018.

Adadi A and Berrada M (2018) Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access 6*, 52138–52160.

Adebayo J, **Muelly M**, **Liccardi I and Kim B** (2020) Debugging tests for model explanations. *Advances in Neural Information Processing Systems 33*, 700–712.

Afzaal M, **Nouri J**, **Zia A**, **Papapetrou P**, **Fors U**, **Wu Y**, **Li X and Weegar R** (2021) Explainable AI for data-driven feedback and intelligent action recommendations to support students self-regulation. *Frontiers in Artificial Intelligence 4*, 723447.

Albreiki B (2022) Framework for automatically suggesting remedial actions to help students at risk based on explainable ML and rule-based models. *International Journal of Educational Technology in Higher Education 19*(1), 1–26.

Arya V, **Bellamy RKE**, **Chen P-Y**, **Dhurandhar A**, **Hind M**, **Hoffman SC**, **Houde S**, **Liao QV**, **Luss R**, **Mojsilović A**, **Mourad S**, **Pedemonte P**, **Raghavendra R**, **Richards J**, **Sattigeri P**, **Shanmugam K**, **Singh M**, **Varshney KR**, **Wei D and Zhang Y** (2019) One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. Preprint, arXiv:1909.03012.

Bach S, **Binder A**, **Montavon G**, **Klauschen F**, **Müller KR and Samek W** (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One 10*(7), 1–46.

Baehrens D, **Schroeter T**, **Harmeling S**, **Kawanabe M**, **Hansen K and Müller K-R** (2010) How to explain individual classification decisions. *The Journal of Machine Learning Research 11*, 1803–1831.

Barocas S, **Selbst AD and Raghavan M** (2020) The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20*. New York, NY: Association for Computing Machinery, pp. 80–89.

Bauman MJ, **Salomon E**, **Walsh J**, **Sullivan R**, **Boxer KS**, **Naveed H**, **Helsby J**, **Schneweis C**, **Lin TY**, **Haynes L**, **Yoder S and Ghani R** (2018) Reducing incarceration through prioritized interventions. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, COMPASS 2018*, volume *18*. New York, NY: Association for Computing Machinery, pp. 1–8, June 2018.

Belle V and Papantonis I (2021) Principles and practice of explainable machine learning. *Frontiers in Big Data 39*, 688969.

Bhatt U, **Andrus M**, **Weller A and Xiang A** (2020a) Machine learning explainability for external stakeholders. Preprint, arXiv:2007.05408.

Bhatt U, **Xiang A**, **Sharma S**, **Weller A**, **Taly A**, **Jia Y**, **Ghosh J**, **Puri R**, **Moura JMF and Eckersley P** (2020b) Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York, NY: Association for Computing Machinery, pp. 648–657.

Boyd S, **Cortes C**, **Mohri M and Radovanovic A** (2012) Accuracy at the top. In *Advances in Neural Information Processing Systems 25*. Red Hook, NY: Curran Associates, Inc., pp. 953–961.

Buçinca Z, **Lin P**, **Gajos KZ and Glassman EL** (2020) Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces, IUI '20*. New York: Association for Computing Machinery, pp. 454–464. https://doi.org/10.1145/3377325.3377498

Cairney P and Oliver K (2017). Evidence-based policymaking is not like evidence-based medicine, so how far should you go to bridge the divide between evidence and policy? *Health research policy and systems*, *15*(1), 1–11.

Carton S, **Helsby J**, **Joseph K**, **Mahmud A**, **Park Y**, **Walsh J**, **Cody C**, **Patterson CPTE**, **Haynes L and Ghani R** (2016) Identifying police officers at risk of adverse events. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: Association for Computing Machinery, pp. 67–76.

Caruana R, **Lou Y**, **Microsoft JG**, **Koch P**, **Sturm M and Elhadad N** (2015) Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: Association for Computing Machinery, pp. 1721–1730.

Chen V, **Li J**, **Kim J S**, **Plumb G, and Talwalkar A** (2022). Interpretable machine learning: Moving from mythos to diagnostics. *Queue*, *19*(6), 28–56.

Chouldechova A, **Putnam-Hornstein E**, **Dworak-Peck S**, **Benavides-Prado D**, **Fialko O**, **Vaithianathan R**, **Friedler SA and Wilson C** (2018) A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Proceedings of Machine Learning Research*, volume *81*. Cambridge, MA: PMLR, pp. 1–15, January 2018.

Coyle D and Weller A (2020) Explaining machine learning reveals policy challenges. *Science 368*(6498), 1433–1434.

Doshi-Velez F and Kim B (2017) Towards a rigorous science of interpretable machine learning. Preprint, arXiv:1702.08608.

Frosst N and Hinton G (2017) Distilling a neural network into a soft decision tree. In *Proceedings of the First International Workshop on Comprehensibility and Explanation in {AI} and {ML} 2017 Co-Located with 16th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017) Conference of the Italian Association*, Aachen, Germany: CEUR–WS. org November 2017.

Gilpin LH, **Bau D**, **Yuan BZ**, **Bajwa A**, **Specter M and Kagal L** (2018) Explaining explanations: An approach to evaluating interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. New York, NY: IEEE, pp. 80–89, May 2018.

Goodman B and Flaxman S (2017) European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine 38*, 50–57.

Guidotti R, **Monreale A**, **Ruggieri S**, **Turini F**, **Giannotti F and Pedreschi D** (2018) A survey of methods for explaining black box models. *ACM Computing Surveys 51*(5), 1–42.

Hase P and Bansal M. (2020). Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5540–5552

Hastie TJ and Tibshirani RJ (1990) *Generalized Additive Models*, volume *43*. Boca Raton, FL: CRC Press.

Hong SR, **Hullman J and Bertini E** (2020) Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction 4*, 1–26.

Hu X, **Rudin C and Seltzer M** (2019) Optimal sparse decision trees. In *Advances in Neural Information Processing Systems 32*. Red Hook, NY: Curran Associates, Inc., pp. 7267–7275, April 2019.

Jacovi A, **Marasović A**, **Miller T and Goldberg Y** (2021) Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*. New York: Association for Computing Machinery, pp. 624–635. https://doi.org/10.1145/3442188.3445923

Jesus S, **Belém C**, **Balayan V**, **Bento J**, **Saleiro P**, **Bizarro P and Gama J** (2021) How can I choose an explainer? An application-grounded evaluation of post-hoc explanations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY: Association for Computing Machinery, pp. 805–815.

Karimi A-H, **Barthe G**, **Balle B and Valera I** (2020) Model-agnostic counterfactual explanations for consequential decisions. In Chiappa S and Calandra R (eds), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume *108* of *Proceedings of Machine Learning Research.* Cambridge, MA: PMLR, pp. 895–905, 26–28 August 2020.

Karimi A-H, **Barthe G**, **Schölkopf B and Valera I** (2021a) A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects. Preprint, arXiv:2010.04050.

Karimi A H, **Schölkopf B, and Valera, I** (2021b). Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 353–362.

Kaufman S, **Rosset S and Perlich C** (2011) Leakage in data mining: Formulation, detection, and avoidance. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery, pp. 556–563.

Kim B, **Khanna R and Koyejo O** (2016) Examples are not enough, learn to criticize! Criticism for interpretability. In *Advances in Neural Information Processing Systems 29*. Red Hook, NY: Curran Associates, Inc., pp. 2280–2288.

Koh PW and Liang P (2017) Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*. Cambridge, MA: JMLR.org, pp. 1885–1894.

König G, **Freiesleben T and Grosse-Wentrup M** (2021) A causal perspective on meaningful and robust algorithmic recourse. Preprint, arXiv:2107.07853.

Lakkaraju H, **Bach SH and Leskovec J** (2016) Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: Association for Computing Machinery, pp. 1675–1684.

Letham B, **Rudin C**, **McCormick TH and Madigan D** (2015) Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics 9*(3), 1350–1371. https://doi.org/10.1214/15-AOAS848

Li B, **Qi P**, **Liu B**, **Di S**, **Liu J**, **Pei J**, **Yi J and Zhou B** (2022) Trustworthy AI: From principles to practices. *ACM Computing Surveys 55*, 1–46.

Lipton ZC (2018) The mythos of model interpretability. *Communications of the ACM 61*, 36–43.

Liu L-P, **Dieterich TG**, **Li N and Zhou Z-H** (2016) Transductive optimization of top k precision. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. Washington, DC: Association for the Advancement of Artificial Intelligence (AAAI) Press, pp. 1781–1787.

Lou Y, **Caruana R and Gehrke J** (2012) Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: Association for Computing Machinery, pp. 150–158.

Lou Y, **Caruana R**, **Gehrke J and Hooker G** (2013) Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: Association for Computing Machinery, pp. 623–631.

Lundberg SM, **Erion G**, **Chen H**, **Degrave A**, **Prutkin JM**, **Nair B**, **Katz R**, **Himmelfarb J**, **Bansal N and Lee S-I** (2020) From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence 2*(1), 56–67.

Lundberg SM, **Erion GG and Lee S-I** (2018a) Consistent individualized feature attribution for tree ensembles. Preprint, arXiv:1802.03888.

Lundberg SM and Lee S-I (2017) A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. Red Hook, NY: Curran Associates, Inc., pp. 4765–4774.

Lundberg SM, **Nair B**, **Vavilala MS**, **Horibe M**, **Eisses MJ**, **Adams T**, **Liston DE**, **Low DKW**, **Newman SF**, **Kim J and Lee SI** (2018b) Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering 2*(10), 749–760.

Molnar C (2019) *Interpretable Machine Learning*. Lulu.com. Available at https://christophm.github.io/interpretable-ml-book/ (Last accessed 14 December 2022).

Mothilal RK, **Sharma A and Tan C** (2020) Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20*. New York, NY: Association for Computing Machinery, pp. 607–617.

Parkhurst J (2016) *The Politics of Evidence: From Evidence-Based Policy to the Good Governance of Evidence*. Oxfordshire, England: Routledge.

Plumb G, **Molitor D and Talwalkar A** (2018) Model agnostic supervised local explanations. In *Advances in Neural Information Processing Systems*. Red Hook, NY: Curran Associates, Inc., pp. 2515–2524.

Potash E, **Ghani R**, **Walsh J**, **Jorgensen E**, **Lohff C**, **Prachand N and Mansour R** (2020) Validation of a machine learning model to predict childhood lead poisoning. *JAMA Network Open 3*(9), e2012734–e2012734.

Poursabzi-Sangdeh F, **Goldstein D**, **Hofman J**, **Vaughan JW and Wallach H** (2021) Manipulating and measuring model interpretability. In *CHI Conference on Human Factors in Computing Systems (CHI '21)*, New york, NY, USA: Association for Computing Machinery (ACM).

Poyiadzi R, **Sokol K**, **Santos-Rodriguez R**, **De Bie T and Flach P** (2020) Face: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY: Association for Computing Machinery, pp. 344–350.

Ramachandran A, **Kumar A**, **Koenig H**, **De Unanue A**, **Sung C**, **Walsh J**, **Schneider J**, **Ghani R and Ridgway JP** (2020) Predictive analytics for retention in care in an urban hiv clinic. *Scientific Reports 10*(1), 1–10.

Ribeiro MT, **Singh S and Guestrin C** (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*. New York, NY: Association for Computing Machinery, pp. 1135–1144.

Ribeiro MT, **Singh S and Guestrin C** (2018) Anchors: High-precision model-agnostic explanations. *AAAI 18*, 1527–1535.

Rodolfa KT, **Salomon E**, **Haynes L**, **Larson J and Ghani R** (2020) Case study: Predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York, NY: Association for Computing Machinery, pp. 142–153.

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence 1*(5), 206–215.

Sajja S, **Aggarwal N**, **Mukherjee S**, **Manglik K**, **Dwivedi S and Raykar V** (2021) Explainable AI based interventions for pre-season decision making in fashion retail. In *8th ACM IKDD CODS and 26th COMAD, CODS COMAD 2021*. New York: Association for Computing Machinery, pp. 281–289.

Saltelli A and Giampietro M (2017) What is wrong with evidence based policy, and how can it be improved? *Futures 91*, 62–71. https://doi.org/10.1016/j.futures.2016.11.012

Samala RK, **Chan H-P**, **Hadjiiski L and Koneru S** (2020) Hazards of data leakage in machine learning: A study on classification of breast cancer using deep neural networks. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume *11314*. Houston, TX: SPIE, pp. 279–284.

Shrikumar A, **Greenside P, and Kundaje A.** (2017). Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3145–3153, August 2017.

Simonyan K, **Vedaldi A and Zisserman A** (2013) Deep inside convolutional networks: Visualising image classification models and saliency maps. Preprint, arXiv:1312.6034.

Sokol K and Flach P (2020) Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20*. New York, NY: Association for Computing Machinery, pp. 56–67.

Tsukimoto H (2000) Extracting rules from trained neural networks. *Transactions on Neural Networks 11*(2), 512–519.

Upadhyay S, **Joshi S, and Lakkaraju H** (2021) Towards robust and reliable algorithmic recourse. In *Advances in Neural Information Processing Systems*, *34*, 16926–16937.

Ustun B and Rudin C (2019) Learning optimized risk scores. *Journal of Machine Learning Research 20*(150), 1–75.

Ustun B, **Spangher A and Liu Y** (2019) Actionable recourse in linear classification. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. New York, NY: Association for Computing Machinery, pp. 10–19.

Ustun B, **Tracà S and Rudin C** (2013) Supersparse linear integer models for interpretable classification. Preprint, arXiv:1306.6677.

van der Waa J, **Robeer M**, **van Diggelen J**, **Brinkhuis MJS and Neerincx MA** (2018) Contrastive explanations with local foil trees. Preprint, arXiv:abs/1806.07470.

Wachter S, **Mittelstadt B and Russell C** (2018) Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Preprint, arXiv:1711.00399.

Weitz K, **Schiller D**, **Schlagowski R**, **Huber T and André E** (2019) "Do you trust me?": Increasing user-trust by integrating virtual agents in explainable AI interaction design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, IVA '19*. New York: Association for Computing Machinery, pp. 7–9. https://doi.org/10.1145/3308532.3329441

Weller A (2019) Transparency: Motivations and challenges. In Samek W, Montavon G, Vedaldi A, Hansen L and Müller KR (eds), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Cham: Springer, pp. 23–40.

Yang H, **Rudin C and Seltzer M** (2017) Scalable Bayesian rule lists. In *Proceedings of the 34th International Conference on Machine Learning—Volume 70, ICML'17*. Cambridge, MA: JMLR.org, pp. 3921–3930.

Ye T, **Johnson R**, **Fu S**, **Copeny J**, **Donnelly B**, **Freeman A**, **Lima M**, **Walsh J and Ghani R** (2019) Using machine learning to help vulnerable tenants in New York City. In *COMPASS 2019—Proceedings of the 2019 Conference on Computing and Sustainable Societies*. New York, NY: Association for Computing Machinery, pp. 248–258, July 2019.

Zahariadis N (2003) *Ambiguity and Choice in Public Policy: Political Decision Making in Modern Democracies*. Washington, D.C.:Georgetown University Press.

Zeiler MD and Fergus R (2014) Visualizing and understanding convolutional networks. In Fleet D, Pajdla T, Schiele B and Tuytelaars T (eds), *European Conference on Computer Vision*. Cham: Springer, pp. 818–833.

Zejnilović L, **Lavado S**, **de Rituerto de Troya ÍM**, **Sim S and Bell A** (2020) Algorithmic long-term unemployment risk assessment in use: Counselors' perceptions and use practices. *Global Perspectives 1*(1), 12908.

Zeng J, **Ustun B and Rudin C** (2017) Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 180*(3), 689–722.