Meaningful Explanations of Black Box AI Decision Systems

Dino Pedreschi,[1] **Fosca Giannotti,**[2] **Riccardo Guidotti,**[2]
Anna Monreale,[1] **Salvatore Ruggieri,**[1] **Franco Turini**[1]
[1]University of Pisa, [2]ISTI-CNR Pisa, Italy
[1]{name.surname}@di.unipi.it, [2]{name.surname}@isti.cnr.it

Abstract

Black box AI systems for automated decision making, often based on machine learning over (big) data, map a user's features into a class or a score without exposing the reasons why. This is problematic not only for lack of transparency, but also for possible biases inherited by the algorithms from human prejudices and collection artifacts hidden in the training data, which may lead to unfair or wrong decisions. We focus on the urgent open challenge of how to construct meaningful explanations of opaque AI/ML systems, introducing the local-to-global framework for black box explanation, articulated along three lines: (i) the language for expressing explanations in terms of logic rules, with statistical and causal interpretation; (ii) the inference of local explanations for revealing the decision rationale for a specific case, by auditing the black box in the vicinity of the target instance; (iii), the bottom-up generalization of many local explanations into simple global ones, with algorithms that optimize for quality and comprehensibility. We argue that the local-first approach opens the door to a wide variety of alternative solutions along different dimensions: a variety of data sources (relational, text, images, etc.), a variety of learning problems (multi-label classification, regression, scoring, ranking), a variety of languages for expressing meaningful explanations, a variety of means to audit a black box.

Open the Black Box

We are evolving, faster than expected, from a time when humans are coding algorithms and carry responsibility of the resulting software quality and correctness, to a time when machines automatically learn algorithms from sufficiently many examples of the algorithms' expected input/output behavior. It is dramatically urgent that machine learning and AI be explainable and comprehensible in human terms; this is instrumental for validating quality and correctness of the resulting systems, and also for aligning the algorithms with human values, as well as preserving human autonomy and awareness in decision making.

On the contrary, the last decade has witnessed the rise of a black box society (Pasquale 2015). Ubiquitous obscure algorithms, increasingly often based on sophisticated machine learning (ML) models trained on (big) data, which predict behavioural traits of individuals, such as credit risk, health status, personality profile. Black boxes map user features into a class or a score without explaining why, because the decision model is not comprehensible to stakeholders, even to expert data scientists. This is worrying not only for the lack of transparency, but also for the possible biases hidden in the algorithms. Machine learning constructs predictive models and decision-making systems based on data, i.e., the digital records of human activities and decisions, such as opinions, movements, preferences, judicial sentences, medical diagnoses, performance scores, etc. Consequently, ML models may reflect human biases and prejudices, as well as collection artifacts and sample selection biases, possibly leading to unfair or simply wrong decisions. Many controversial cases have already highlighted that delegating decision-making to black box algorithms is critical in many sensitive domains, including crime prediction, personality scoring, image classification, personal assistance, and more (Pedreschi et al. 2018).

A missing step in the construction of an ML model is precisely the explanation of its logic, expressed in a comprehensible, human-readable format, that highlights the biases learned by the model, allowing to understand and validate its decision rationale. This limitation impacts not only information ethics, but also accountability, safety and industrial liability (Danks and London 2017; Kingston 2016; Kroll et al. 2017). Companies increasingly market services and products with embedded ML components, often in safety-critical industries such as self-driving cars, robotic assistants, domotic IoT systems, and personalized medicine. An inherent risk of these components is the possibility of inadvertently making wrong decisions, learned from artifacts or spurious correlations in the training data, such as recognizing an object in a picture by the properties of the background, due to a systematic bias in training data collection. How can companies trust their products without understanding the rationale of their machine learning components?

Likewise, the use of machine learning models in scientific research, for example in medicine, biology, socio-economic sciences, requires explanations not only for trust and acceptance of results, but also for the very sake of the openness of scientific discovery and the progress of research.

An explanation technology would be of immense help to companies for creating safer, more trustable products, and better managing any possible liability they may have.

From the citizen's perspective, the EU General Data Protection Regulation (GDPR), entered into force in Europe on 25 May 2018, introduces a right of explanation for individuals to obtain "meaningful information of the logic involved" when automated decision making takes place with "legal effects" on individuals "or similarly significantly affecting" them[1]. Without an enabling technology for explanation this right will either remain "dead letter", or will just outlaw many opaque AI systems (Goodman and Flaxman 2016; Malgieri and Comandé 2017). *Explanation is at the heart of a responsible, human-centric AI, across multiple industry sectors and scientific disciplines. An inescapable challenge, a cornerstone to develop AI systems aimed at empowering and engaging people, not at replacing them.*

Despite the soaring recent body of research on interpretable ML a practical, widely applicable technology for explainable AI has not emerged yet. The challenge is hard, as explanations should be sound and complete in statistical and causal terms, and yet comprehensible to multiple stakeholders such as the users subject to the decisions, the developers of the automated decision system, researchers, data scientists and policy makers, authorities and auditors, including regulation and competition commissions, civil rights societies, etc. Stakeholders should be empowered to reason on explanations, to understand how the automated decision-making system works on the basis of the inputs provided by the user; what are the most critical features; whether the system adopts latent features; how a specific decision is taken and on the basis of what rationale/reasons; how the user could get a better decision in the future.

The problem can be articulated in two different flavors:

- **eXplanation by Design** (*XbD*): given a dataset of training decision records, how to develop a machine learning decision model *together with* its explanation;
- **Black Box eXplanation** (*BBX*): given the decision records produced by an obscure black box decision model, how to reconstruct an explanation for it.

In the *XbD* problem setting, we would like to empower the data scientist in charge of developing a decision ML model with the means to provide also an explanation of the model's logic, in order to prevent from making unfair, inaccurate or simply wrong decisions learned from artifacts and biases hidden in the training data and/or amplified or introduced by the learning algorithm. At the same time, we would like to preserve the liberty of the data scientist to use any kind of ML task, including non-interpretable models such as complex deep learning or ensemble models. In this scenario, where the data scientist has full control over the model's creation process, the development of an explanation is essentially a further validation step in assessing the quality of the output model (in addition to testing for accuracy, absence of overfitting, etc.). The explanation is also an extra deliverable of the learning process, sustaining transparency and the trust of the stakeholders who will adopt the model.

In the harder *BBX* problem setting, we would like to empower the data scientist with means for auditing and finding an explanation for a black box designed by others. In this case, the original dataset on which the black box was trained is not known, and neither are the internals of the model. In fact, only the decision behaviour of the black box can be observed. In our framework, we assume that the black box can be queried to acquire data about its decision behaviour, or that such data can be gathered by participating individuals.

We focus on the open challenge of how to construct meaningful explanations in the XbD and BBX cases, and delineate a novel research direction inspired by early methods for local explanations (including our own), i.e., methods to explain why a certain specific case has received its own classification outcome. We propose a new ***local-first explanation framework***: expressive logic rule languages for inferring local explanations, together with bottom-up generalization algorithms to aggregate an exhaustive collection of local explanations into a global one, optimizing jointly for simplicity and fidelity in mimicking the black box. We argue that the local-first approach has the potential to advance the state of art significantly, opening the door to a wide variety of alternative technical solutions along different dimensions: the variety of data sources (relational, text, images, etc.), the variety of learning problems (binary classification, multi-label classification, regression, scoring, ranking, etc.), the variety of languages for expressing meaningful explanations, the variety of means to audit the black box.

Ideally, more informative ***causal explanations*** should be provided, that capture the causal relationships among the (endogenous as well as exogenous) variables and the decision, based on the data observed by appropriately querying the black box. Why moving from purely statistical to causal explanations? ML models are used to classify under the assumption of independent and identically distributed data, generated by the same fixed distribution. Causal models would enable to classify under changing distributions, e.g., would allow to perform what-if reasoning under realistic dynamic scenarios.

A Very Succinct State of the Art

Although attempts to tackle *interpretable machine learning* and *discrimination-aware data mining* exist for several years now, there has been an exceptional growth of research efforts in the last couple of years, with new emerging keywords such as *black box explanation* and *explainable AI*. We refer to our comprehensive, up-to-date survey (Guidotti et al. 2018b), and account briefly here for the major recent trends. An early study on the nature of explanations from a psychological viewpoint is (Leake 1992). Many approaches to the XbD problem attempt at explaining the *global* logic of a black box by an associated interpretable classifier that mimics the black box. These methods are mostly designed for specific machine learning models, i.e., they are not agnostic, and often the interpretable classifier consists in a decision tree or in a set of decision rules. For example, decision trees have been adopted to explain neural networks (Krishnan, Sivakumar, and Bhattacharya 1999) and tree ensembles (Tan, Hooker, and Wells 2016), while decision rules have been widely used to explain neural networks (Augasta and Kathirvalavakumar 2012;

[1] http://ec.europa.eu/justice/data-protection/

Andrews, Diederich, and Tickle 1995) and SVM (Support Vector Machines) (Fung, Sandilya, and Rao 2005). A few methods for global explanation are agnostic w.r.t. the learning model (Lou, Caruana, and Gehrke 2012; Henelius et al. 2014).

A different stream of approaches, still in the *XbD* setting, focuses on the *local* behavior of a black box (Guidotti et al. 2018b), searching for an explanation of the decision made for a specific instance. Some such approaches are model-dependent and aim, e.g., at explaining the decisions of neural networks by means of saliency masks, i.e., the portions of the input record (such as the regions of an image) that are mainly responsible for the outcome (Xu et al. 2015; Zhou et al. 2016; Nugent and Cunningham 2005). A few more recent methods are model-agnostic, such as LIME (Ribeiro, Singh, and Guestrin 2016; Singh and Anand 2018). The main idea is to derive a local explanation for a decision outcome y on a specific instance x by learning an interpretable model from a randomly generated neighborhood of x, where each instance in the neighborhood is labelled by querying the black box. An extension of LIME using decision rules (called Anchors) is presented in (Ribeiro, Singh, and Guestrin 2018), which uses a bandit algorithm that randomly constructs the rules with the highest coverage and precision. Our group has designed LORE (Guidotti et al. 2018a), a local explanator that builds a focused exploration around the target point, and delivers explanations in the form of highly expressive rules together with *counterfactuals*, suggesting the changes in the instance's features that would lead to a different outcome. When the training set is available, decision rules are also widely used to proxy a black box model by directly designing a transparent classifier (Guidotti et al. 2018b) which is locally or globally interpretable on its own (Lakkaraju, Bach, and Leskovec 2016; Malioutov et al. 2017; Craven and Shavlik 1995).

To sum up, despite the soaring attention to the topic, the state of the art to date still exhibits ad-hoc, scattered results, mostly hard-wired with specific models. A widely applicable, systematic approach with a real impact has not emerged yet. This is a tremendous obstacle to develop a human-centric AI, a danger for the digital society. In our view, a black box explanation framework should be:

1. **model-agnostic**: it can be applied to any black box model;
2. **logic-based**: so that explanations can be made comprehensible to humans with diverse expertise, and support their reasoning, and be extensible to handle causal reasoning by using meta-logic and abduction;
3. **both local and global**: it can explain both individual cases and the overall logic of the black-box model;
4. **high-fidelity**: it provides a reliable and accurate approximation of the black box behavior.

The four desiderata do not coexist in current proposals. Logic-based decision rules have proven useful in the sub-problem of explaining discrimination from a purely data-driven perspective, as demonstrated in the lively stream of research in discrimination-aware data mining, started by our research group in 2008 (Pedreschi, Ruggieri, and Turini 2008; Ruggieri, Pedreschi, and Turini 2010), but it is unlikely that rules in their simplest form will solve the general explanation problem. Global rule-based models, trained on black box decision records, are often either inaccurate, over-simplistic proxies of the black box, or too complex, thus compromising interpretability. On the other hand, purely local models, such as LIME or our method LORE mentioned above, do not yield an overall proxy of the black box, hence cannot solve the XbD and BBX problems in general terms.

How to Construct Meaningful Explanations?

To tackle the challenge, we propose a broad direction of research for constructing meaningful explanations: a *local-to-global framework for explanation-by-design* that, beyond statistical explanations, also investigates *causal explanations*. Alongside, it is also needed to develop: *(i)* an *explanation infrastructure* for the benchmarking of the methods, equipped with platforms for the users' assessment of the explanations and the crowd-sensing of observational decision data; *(ii)* an *ethical-legal framework*, both for compliance and impact of the developed methods on current legal standards; and *(iii)* a wide-variety of case studies of explanation-by-design in challenging domains, e.g., in *health* and *fraud detection* applications, to validate the approaches.

Local-to-Global for Explanation by Design

Let us consider the *XbD* problem of discovering an explanation for a high-quality black box model b learned over a training dataset of labelled examples (x, y), where y is the class label and x is a vector of observed features; let us concentrate on binary classification, i.e., $y \in \{0, 1\}$. Our framework works under the following three assumptions.

H1: Logic explanations. The cognitive vehicle for offering explanations should be as close as possible to the language of reasoning, that is *logic*. From simple propositional rules up to more expressive, possibly causal and counterfactual logic rules, many options of varying expressiveness exist to explore the trade-off between accuracy and interpretability of explanations.

H2: Local explanations. The decision boundary for the black box b can be arbitrarily complex over the training set, but *in the neighborhood of each specific data point* (x, y) *there is a high chance that the decision boundary is clear and simple*, likely to be accurately approximated by an interpretable explanation.

H3: Explanation composition. There is a high chance that *similar data points admit similar explanations*, and similar explanations are likely to be composed together into slightly more general explanations.

H2 and H3 are motivated by the observation that if all data points in the training set are surrounded by complex decision boundaries, or if any two data points admit very different explanations due to different decision boundaries, then b is likely to be in overfitting, unable to generalize from training data of insufficient quality, thus contradicting the basic assumption. These assumptions suggest a two-step, *local-first* approach to the *XbD* problem, also applicable to the *BBX* problem if the black box can be queried without limitations:

Local Step. For any example in the training set (or any other example) labeled by the black box b, i.e., for any specific (x, y'), where $y' = b(x)$ is the label assigned by b to x, query b to label a sufficient set of examples (*local dataset*) in the neighborhood of (x, y'), which are then used to derive an explanation $e(x, y')$ for x. The explanation answers the question: why b has assigned class y' to x? and, possibly, also its *counterfactual*: what should be changed in x to revert the outcome? (see our recent paper (Guidotti et al. 2018a).)

Local-to-Global Step. Consider as an initial global explanation the set of all local explanations $e(x, y')$ constructed at the local step for each available individual example (x, y'), and synthesize a smaller set by iteratively composing and generalizing together similar explanations, optimizing for simplicity and fidelity.

The local-first explanation framework may be articulated along different dimensions: the *variety of data sources* (relational, text, images, ...), the *variety of learning problems* (binary classification, multi-label classification, regression, scoring, ranking, ...), the *variety of languages for expressing meaningful explanations*. The various technical options for each dimension yield a large number of combinations that call for suitable explanation models and algorithms. Local explanation methods need to consider different ways for auditing the black box, as well as many alternatives for learning local explanations. The local-to-global methods need to consider alternative ways of synthesizing multiple explanations into more general ones.

Regarding *reasoning mechanisms*, simply providing the user with the explanations computed by the algorithms may not work. There may be too many local explanations, or the user may need to ask specific questions: *For what reasons are the applications of a specific population or profile rejected? Which explanations highlight potential discrimination of, e.g., protected minorities? What combinations of features are most strongly correlated to (or are a cause of) a negative decision?* It is needed to design reasoning mechanisms capable of mapping such high-level questions into queries to the global explanation, providing answers for the users. Such reasoning methods can be realized in logic by using *meta-logic*. Moreover, appropriate interfaces need to be designed, to convey the answers in meaningful visual and textual forms, as well as visual exploration for advanced users, leveraging also on available visual analytics tools, e.g., (Krause, Perer, and Ng 2016).

From Statistical to Causal Explanations

Machine learning leverages statistical associations in observational data that, in general, do not convey information about the causal dependencies among the observed variables and the unobserved confounding variables. Nonetheless, ML models are often used within decision making processes in the real world, so that certain observed features are the causes of specific effects, i.e., the decision outcomes. The science of causal inference and learning has developed tools, such as Causal Graphical Models and Structural Causal Models (Peters, Janzing, and Scholkopf 2017; Pearl and MacKenzie 2018; Bareinboim and Pearl 2016), to answer certain *interventional* questions (what if something changes?) or *retrospective* questions (what would have happened had a different choice been made?). Such tools have not been applied specifically to explain ML models, with a few exceptions dealing on specific issues, such as discrimination inference (Zhang and Bareinboim 2018; Bonchi et al. 2017; Caliskan, Bryson, and Narayanan 2017; Caravagna et al. 2016) and reasoning on sample selection bias (Bareinboim and Pearl 2016). It is natural to investigate the causal structure that an ML model has inferred from the training data, establishing a **link between causality and black box explanation**. First, it is interesting to explore how the techniques for data-driven causal discovery, aimed at reconstructing plausible graphs of causal dependencies in observational data (see, e.g., (Huang et al. 2018)), may be used to achieve more informative and robust explanations. Second, it is promising to explore how the techniques for causal inference may be used for driving the audit of a black box in the local explanation discovery.

A Platform for Explainable AI

It is crucial to establish an "open science" infrastructure for sharing experimental data and explanation algorithms with the research community, creating a common ground for researchers working on explanation of black boxes from different domains. It is also crucial to develop dedicated participatory platforms, which support the engagement of a crowd of users to check the comprehensibility and usefulness of the provided explanations for the decision they got, thus supporting campaigns of validation of the proposed technical solutions, and to provide data about their experience in interacting with black-box services.

The platform should enable validating our approach in realistic cases in challenging domains, such as health-care, e.g., explaining systems such as DoctorAI, based on a multi-label Recurrent Neural Network trained on patients' Electronic Health Records (Choi et al. 2016). The objective is to devise explanations for "extreme" multi-label classification, also delivering causal explanations, based on curated data sources such as MIMIC-III (Johnson et al. 2016). Other example domains include fiscal fraud detection (Bonchi et al. 1999) and car insurance telematics and driver profiling (Nanni et al. 2016)).

An Ethical/Legal Framework for Explanation

The research direction illustrated in this paper has a strong ethical motivation. It aims to empower users against undesired, possibly illegal, effects of black-box automated decision-making systems which may harm them, exploit their vulnerabilities, and violate their rights and freedom. On one hand, it is needed to comply with the relevant legal frameworks, the European GDPR in particular. On the other hand, it is needed to investigate the impacts towards ethics and law. In particular, are the provided explanations useful *(i)* for the realization of the *right of explanation* provisions of the GDPR? *(ii)* for the industrial development of AI-powered services and products? *(iii)* in revealing new forms of discrimination towards new vulnerable social groups?

Acknowledgement. This work is partially supported by the European Community's H2020 Program under the funding scheme "INFRAIA-1-2014-2015: Research Infrastructures" G.A. 654024, http://www.sobigdata.eu, "SoBigData: Social Mining & Big Data Ecosystem".

References

Andrews, R.; Diederich, J.; and Tickle, A. B. 1995. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowl.-Based Syst.* 8(6):373–389.

Augasta, M. G., and Kathirvalavakumar, T. 2012. Reverse engineering the neural networks for rule extraction in classification problems. *NPL* 35(2):131–150.

Bareinboim, E., and Pearl, J. 2016. Causal inference and the data-fusion problem. *PNAS* 113(27):7345–7352.

Bonchi, F.; Giannotti, F.; Mainetto, G.; and Pedreschi, D. 1999. A classification-based methodology for planning audit strategies in fraud detection. In *KDD*, 175–184. ACM.

Bonchi, F.; Hajian, S.; Mishra, B.; and Ramazzotti, D. 2017. Exposing the probabilistic causal structure of discrimination. *IJDSA* 3(1):1–21.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356:183–186.

Caravagna, G.; Graudenzi, A.; Ramazzotti, D.; Sanz-Pamplona, R.; De Sano, L.; Mauri, G.; Moreno, V.; Antoniotti, M.; and Mishra, B. 2016. Algorithmic methods to infer the evolutionary trajectories in cancer progression. *PNAS* 113(28):E4025–E4034.

Choi, E.; Bahadori, M. T.; Schuetz, A.; Stewart, W. F.; and Sun, J. 2016. Doctor AI: Predicting clinical events via recurrent neural networks. In *PMLR*, 301–318.

Craven, M., and Shavlik, J. W. 1995. Extracting tree-structured representations of trained networks. In *NIPS*, 24–30.

Danks, D., and London, A. J. 2017. Regulating autonomous systems: Beyond standards. *IEEE IS* 32(1):88–91.

Fung, G.; Sandilya, S.; and Rao, R. B. 2005. Rule extraction from linear support vector machines. In *KDD*, 32–40. ACM.

Goodman, B., and Flaxman, S. 2016. EU regulations on algorithmic decision-making and a "right to explanation". In *ICML*.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Pedreschi, D.; Turini, F.; and Giannotti, F. 2018a. Local rule-based explanations of black box decision systems. *arXiv:1805.10820*.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018b. A survey of methods for explaining black box models. *CSUR* 51(5):93:1–93:42.

Henelius, A.; Puolamäki, K.; Boström, H.; Asker, L.; and Papapetrou, P. 2014. A peek into the black box: exploring classifiers by randomization. *DAMI* 28(5-6):1503–1529.

Huang, B.; Zhang, K.; Lin, Y.; Scholkopf, B.; and Glymour, C. 2018. Generalized score functions for causal discovery. In *KDD*, 1551–1560. ACM.

Johnson, A. E. W.; Pollard, T. J.; Shen, L.; Lehman, L. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3:160035.

Kingston, J. K. C. 2016. Artificial intelligence and legal liability. In *SGAI Conf.*, 269–279. Springer.

Krause, J.; Perer, A.; and Ng, K. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *CHI*, 5686–5697. New York, NY, USA: ACM.

Krishnan, R.; Sivakumar, G.; and Bhattacharya, P. 1999. Extracting decision trees from trained neural networks. *Pattern recognition* 32(12).

Kroll, J. A.; Huey, J.; Barocas, S.; Felten, E. W.; Reidenberg, J. R.; Robinson, D. G.; and Yu, H. 2017. Accountable algorithms. *U. of Penn. Law Review* 165:633–705.

Lakkaraju, H.; Bach, S. H.; and Leskovec, J. 2016. Interpretable decision sets: A joint framework for description and prediction. In *KDD*, 1675–1684. ACM.

Leake, D. B. 1992. *Evaluating explanations: A content theory*. Lawrence Erlbaum Associates.

Lou, Y.; Caruana, R.; and Gehrke, J. 2012. Intelligible models for classification and regression. In *KDD*, 150–158. ACM.

Malgieri, G., and Comandé, G. 2017. Why a right to legibility of automated decision-making exists in the general data protection regulation. *IDPL* 7(4):243–265.

Malioutov, D.; Varshney, K. R.; Emad, A.; and Dash, S. 2017. Learning interpretable classification rules with boolean compressed sensing. In *Transparent Data Mining for Big and Small Data*. Springer. 95–121.

Nanni, M.; Trasarti, R.; Monreale, A.; Grossi, V.; and Pedreschi, D. 2016. Driving profiles computation and monitoring for car insurance CRM. *ACM Trans. Intell. Syst. Technol.* 8(1):14:1–14:26.

Nugent, C., and Cunningham, P. 2005. A case-based explanation system for black-box systems. *AIR* 24(2):163–178.

Pasquale, F. 2015. *The black box society: The secret algorithms that control money and information*. HUP.

Pearl, J., and MacKenzie, D. 2018. *The Book of Why: the new science of cause and effect*. Basic Books.

Pedreschi, D.; Giannotti, F.; Guidotti, R.; Monreale, A.; Pappalardo, L.; Ruggieri, S.; and Turini, F. 2018. Open the black box data-driven explanation of black box decision systems. *arXiv:1806.09936*.

Pedreschi, D.; Ruggieri, S.; and Turini, F. 2008. Discrimination-aware data mining. In *KDD*, 560. ACM.

Peters, J.; Janzing, D.; and Scholkopf, B. 2017. *Elements of causal inference: foundations and learning algorithms*. MIT Press.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *KDD*, 1135–1144. ACM.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-precision model-agnostic explanations. In *AAAI*, 1527–1535.

Ruggieri, S.; Pedreschi, D.; and Turini, F. 2010. Data mining for discrimination discovery. *TKDD* 4(2):9:1–9:40.

Singh, J., and Anand, A. 2018. Exs: Explainable search using local model agnostic interpretability. *arXiv:1809.03857*.

Tan, H. F.; Hooker, G.; and Wells, M. T. 2016. Tree space prototypes: Another look at making tree ensembles interpretable. *arXiv:1611.07115*.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2048.

Zhang, J., and Bareinboim, E. 2018. Fairness in decision-making: The causal explanation formula. In *AAAI*, 2037.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *CVPR*, 2921–2929. IEEE.