Interpretable ML model for quality control of locks using counterfactual explanations

1st Tim Andersson
School of Innovation, Design and Engineering
Mälardalen University
Västerås, Sweden
Tim.Andersson@mdu.se

2nd Markus Bohlin
School of Innovation, Design and Engineering
Mälardalen University
Västerås, Sweden
Markus. Bohlin @mdu.se

3rd Mats Ahlskog
School of Innovation, Design and Engineering
Mälardalen University
Västerås, Sweden
Mats. Ahlskog @mdu.se

4th Tomas Olsson
School of Innovation, Design and Engineering
Mälardalen University
Västerås, Sweden
Tomas.Olsson@ri.se

***Abstract*—This paper presents an interpretable machine-learning model for anomaly detection in door locks using torque data. The model aims to replace the human tactile sense in the quality control process, reducing repetitive tasks and improving reliability. The model achieved an accuracy of 96%, however, to gain social acceptance and operators' trust, interpretability of the model is crucial. The purpose of this study was to evaluate an approach that can improve interpretability of anomalous classifications obtained from an anomaly detection model. We evaluate four instance-based counterfactual explanators, three of which, employ optimization techniques and one uses, a less complex, weighted nearest neighbor approach, which serve as our baseline. The former approaches, leverage a latent representation of the data, using a weighted principal component analysis, improving plausibility of the counter factual explanations and reduces computational cost. The explanations are presented together with the 5-50-95th percentile range of the training data, acting as a frame of reference to improve interpretability. All approaches successfully presented valid and plausible counterfactual explanations. However, instance-based approaches employing optimization techniques yielded explanations with greater similarity to the observations and was therefore concluded to be preferable despite the higher execution times (4-16s) compared to the baseline approach (0.1s). The findings of this study hold significant value for the lock industry and can potentially be extended to other industrial settings using timeseries data, serving as a valuable point of departure for further research.**

Keywords—Explainable artificial intelligence, Counterfactual explanation, Anomaly detection, Principal component analysis

I. Introduction

Machine Learning (ML) has emerged as a valuable tool to handle complex, high-dimensional data and has often outperformed traditional statistical and rule-based approaches, resulting in improved accuracy [1]. When building a machine learning model for anomaly detection, it's typical to use an unsupervised learner [2]. This means that the training dataset lacks labelled data or a one-class learner whereas the training dataset comprises only one class [3]. This is due to the scarcity of observations of anomalous behavior compared to normal behavior. Additionally, anomalous behaviors can vary significantly which makes it difficult to obtain a representative sample [4]. There are several algorithms commonly used to create an anomaly detection model, including One-Class Support Vector Machine (OCSVM) and One-class K-Nearest Neighbor (OCKNN), among others, which have shown promising results in various applications [2], [5]. Unfortunately, the use of ML models can render the decision-making process less transparent due to the complex, non-linear relationships that they are capable of learning, making them a blackbox model, which is difficult for humans to comprehend. An interpretable ML model is necessary to ensure social acceptance and make the model more trustworthy for humans, as it allows for causal relationships to be evaluated and information to be extracted to partially explain the model's decisions [6]. In the context of interpretable ML, it is crucial to establish the intended purpose of the model's decision explanation. This ensures that the explanation is appropriate for the target audience and effectively conveys the desired information. Depending on the specific context, the explanation may require a complete causal attribution of all influencing factors, such as in legal cases and model debugging. Alternatively, it may require a simplified, human-friendly explanation, which is the focus of this study. It has been shown that humans often favor explanations that are contrastive since they are easy to understand [7]. So, instead of explaining how each factor of an observation contributes to a specific decision from an ML model, it can be a better approach to show an observation with similar characteristics that the model made a different decision on i.e., counterfactual explanation [6], [8], [9], which is the approach considered in this study. In a recent study [10], a ML approach was evaluated to replace the human tactile sense in the task of quality control of door locks. Although historically only human operators were trusted to perform this task, an automated approach is under development to obtain a more objective solution and alleviate this repetitive task from the operators. The ML models outperformed the current method of using the human tactile in detecting faulty locks but, without providing any explanations. The purpose of this study was to evaluate an approach to improve interpretability of an anomalous classification obtained from a ML model used to detect mechanical anomalies in door

978-1-6654-0932-2/22/$31.00 ©2022 IEEE

locks. The model was trained on data from a torque sensor that measures the force required to turn the lock 360 degrees, similar to how humans perform the task using their tactile sense. The research questions for this study are:

- Which method is suitable to use in the context of anomaly detection using torque sensor data to ensure the interpretability of the model's result and what measures should be used to compare and evaluate different approaches?

II. Reviewed literature

A counterfactual explanation can be created in numerous ways, such as, utilize an instance-based strategy using a distance metric to retrieve a similar observation from a known dataset or created synthetically by specific feature permutations using optimization techniques to minimize a cost function, which is particularly common in literature [8]. Various criteria have been proposed to evaluate the efficacy of counterfactual explanations, but some frequently mentioned are *Validity*, which means that the explanation actually changes the model's decision, *Actionability*, the permutated features must be allowed to be changed in reality (in this study, all features can be changed), *Plausibility*, the explanation should be coherent with verified observations in a dataset which encountered the same decision from the model, *Similarity/Proximity*, the explanation should be close to the original observation with respect to some distance metric, *Sparsity*, minimal permutation to as few features as possible. The methods applied to create an explanation and to meet a specific set of criteria are often experimentally adapted with respect to the nature of the dataset, the intended user and the application-specific needs, hence, there is no approach that fits all scenarios [8], [11], [12].

One of the most known approaches for generating counterfactual explanations conceptualizes the task as an optimization problem, wherein a cost function is employed to effectuate a weighted summation of the desired model output and the distance between the counterfactual instance and the original observation [9]. Using an optimization approach, one can elaborate with different weights and terms in the cost function to get desired properties of the explanations that best fits a specific application. However, the use of optimization methods may incur substantial computational cost for high-dimensional datasets, as in this study, where we have 3600 features per observation. Certain works in the literature have leveraged case-based techniques as a prelude to the optimization process, potentially accelerating convergence and enhancing both plausibility and validity of the generated explanations [13], [14], [15]. Another innovative approach utilizes latent data representations to mitigate computational costs further, while also potentially improving plausibility, as in [16]. Despite the advancements in counterfactual explanations in general, there remains an evident gap in the literature concerning their application in anomaly detection models utilizing torque sensor data. This presents an unexplored avenue for both academic and industrial research, making the current study a valuable contribution.

III. Method

In this section, a detailed description of the main steps performed in this study namely data collection and preprocessing, model training and counterfactual explanations.

A. Data collection, pre-processing and model training

The first step was the data collection. Two sets of locks were created, a functional set, and an anomalous set that consisted of locks that had been contaminated with dust and sand, similar to what can happen in the factory. The contamination can result in an unsmooth jerky motion when the locks are operated and can be detected using a torque sensor. In this study, we used the same equipment as in [10], where an electric motor was used to rotate the locks at a constant angular velocity of 30 degrees per second. The torque applied to turn each lock was measured using a sensor with a range of 0.1-200 Nm and a sensitivity of ±0.02 Nm. The measurements were then mapped to a specific angular position with a resolution of 0.1 degrees, resulting in 3600 features for each lock for a complete 360-degree rotation. In total, one dataset of 143 fully functional locks (non-anomalous observations) of size 143x3600 and another dataset of 36 anomalous observations of size 36x3600 were obtained. When rotating the lock, the required torque for maintaining a constant angular velocity fluctuates throughout a complete revolution due to variations in the internal friction at different angular positions. This variability can lead to abrupt shifts in the lock's angular position caused by a mechanical spring effect and measurement equipment tolerances. The computer's sampling rate of 1ms was not enough to cope with these sudden shifts, resulting in missed measurements ranging from 0-5% for specific angular positions. To address this issue, a linear interpolation between adjacent data points was applied. To reduce noise the data were processed using a mean moving-window approach of size 10. To get equal weight from each feature they need to have a similar range, this was solved by normalizing each feature based on the training data such that the mean was zero and the standard deviation was one (z-score normalization). The same normalization parameters were used to normalize the anomalous observations as well. Next, we trained and evaluated an OCSVM-model with the normalized data using a Monte Carlo approach. As OCSVM specializes in anomaly detection, the model was exclusively trained on the non-anomalous dataset however, for validation, we incorporated both non-anomalous and anomalous observations. We randomly split the non-anomalous dataset into 114 training observations and 29 validation observations. Additionally, all 36 anomalous observations were included in the validation set. This process was replicated 100 times to compute an averaged Receiver Operating Characteristic (ROC) curve and determine the mean accuracy. We chose the OCSVM model since it's one of the commonly used models for anomaly detection [2], [5] and has been proven efficient for anomaly detection in locks [10].

B. Counterfactual explanations

Drawing inspiration from prior work [13], [14], [15], our research undertakes a detailed comparative assessment of four instance-based counterfactual explanators. Among these, three incorporate optimization techniques after retrieving the Weighted Nearest Neighbor (WNN), while the fourth utilizes a straightforward WNN alone. This latter approach stands as our

baseline. The weights are inversely proportional to the absolute deviation from the mean of the training dataset. These specific weights are crafted to identify a non-anomalous neighbor that aligns closely with anomaly features unlikely to be the root cause of the anomalous classification, thereby offering a nuanced, contrastive explanation. The rationale for preferring optimization-techniques over WNN, lies in their potential to yield greater similarity and generate explanations that closely align with the model's decision boundary. Consequently, they hold the promise of uncovering the model's vulnerabilities. Nevertheless, it is essential to acknowledge that optimization techniques can entail heightened computational costs. To address this, we evaluate three distinct optimization techniques characterized by varying levels of complexity.

The optimization techniques leverage insights derived from instances in the training data to guide the optimization process, thereby enhancing the plausibility and validity of the generated explanations. The first approach combines a Genetic Algorithm (GA) with Sequential Quadratic Programming (SQP) [17]. Initially, GA is applied, followed by SQP, with a search space bounded by the closest non-anomalous (p) and anomalous observations (q). Using a constrained search space in the proximity of the training data and the anomalous observation facilitates plausibility, similarity and lower computational cost. We employ a population size of 10 and set a maximum of 10 generations, although this limit was not reached in our experiments. All generated populations adhere to predefined search space, and default values are retained for other parameters. The second approach employs SQP with default settings [18], employing the same bounded search space as the first approach. The third approach also leverages SQP with default settings, albeit constrained along a linear path between the nearest non-anomalous and anomalous observations. In both the second and third approaches, the initial starting point is set as the closest non-anomalous point. Notably, all three approaches are subjected to an inequality constraint linked to the classification score $y(x)$, obtained from the OCSVM model. This constraint guarantees that the resulting counterfactual explanation consistently maintains validity by being classified as a non-anomalous observation. Specifically, the OCSVM model assigns a score below zero to anomalous observations and above zero to non-anomalous ones. The overarching objective across all approaches is the minimization of the distance d between the two points. We elaborated with different distance metrics such as, Euclidean and Manhattan-distance which are commonly used but, the cosine distance yielded overall more stable performance and is therefore the distance metric used in this study. The following constraints are expressed in standard null form:

$$g1(\boldsymbol{x}) = -y(\boldsymbol{x}) < 0 \quad (1)$$

$$g2(\boldsymbol{x}) = [x_i, x_{i+1}, ..] < [max(p_i, q_i), max(p_{i+1}, q_{i+1}), ...] \quad (2)$$

$$g3(\boldsymbol{x}) = -[x_i, x_{i+1}, ..] < [min(p_i, q_i), min(p_{i+1}, q_{i+1}), ...] \quad (3)$$

The initial step in all three optimization-techniques involves the transformation of the data into a latent representation using a weighted Principal Component Analysis (PCA). A dataset was created by producing multiple instances of the same anomaly which were combined with the training data to create a single balanced dataset before utilizing the PCA algorithm, this enhances the discriminative power of the latent representation. The number of components to retain was a compromise between execution time and resolution of the transformation which in this study was selected to 10 components.

A latent representation alleviates computational complexity during the optimization and serves to mitigate the potential for generating unrealistic permutations, consequently enhancing the plausibility and similarity. The feature weights used with PCA, based on the absolute difference between the training data mean and anomalies, enhances causal interpretability by emphasizing influential features. This reduces the ability of permuting nearby features, making explanations more causally informative, as features significantly deviating from the mean are more likely to have a greater role in classifying an observation as an anomaly. While this study prioritizes plausibility over sparsity, the use of PCA tend to align features with lower variance more closely with the observation it started from instead of the anomaly. This adversely impact the overall sparsity of the counterfactual explanation. To assess the uncertainty of the latent representation of the data resulting from PCA, we used 100 bootstrapped means to calculate a 5-95th percentile confidence interval for each observation in the latent space. The observation corresponding to the 95th percentile confidence interval width was selected and transformed back to the original feature space to implicitly illustrate the uncertainty of the latent representation, see Fig. 1. The average width of the confidence interval was 26Nmm, which is similar to the sensitivity of the sensor used for data collection (20Nmm). We chose to present the counterfactual explanations as torque curves since the operators already know that variations in torque amplitude are directly linked to the change of needed force to turn the lock. This makes it easy for the operators to verify the model's decision if needed using their tactile sense.

All counterfactual explanations are presented alongside the training dataset's median and the 5-95th percentile range, offering contextual information, with the intension of improving interpretability and trustworthiness. The counterfactual explanations are subjectively evaluated by an expert in torque measurements from these locks, assessing the plausibility and similarity visually, prioritizing human-friendly explanation. Additionally, we also compare the median execution time between the four approaches.

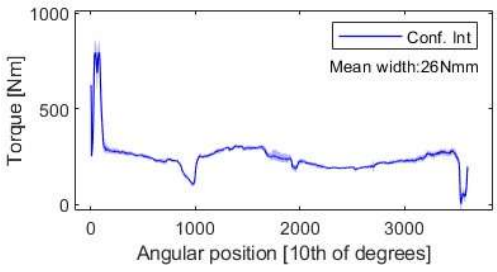

Fig. 1. Showcases the bootstrapped mean confidence interval for the latent representation derived from the PCA transformation, implicitly visualized within the feature space.

IV. Results

In Tab. 1, it can be observed that the line bounded SQP was 61-74% faster than the other two optimization approaches although the WNN was 98-99% faster than all the others. In Fig. 2, we present the performance evaluation of the OCSVM model. The OCSVM model utilized in our study, has a mean accuracy of 96%. It also accurately classifies anomalous observations 100% correctly.

We selected 6 anomalous observations of locks at random from a total of 36 for presentation in Fig. 3. While all 36 locks underwent identical evaluations with congruent outcomes, a subset of 6 observations sufficiently demonstrates the results. Each plot in Fig. 3, showcases the standardized cosine distances, computed in the feature space, for each counterfactual approach, offering a quantifiable measure to discern the similarity differences across methods. These distances have been standardized relative to the outcome of the WNN approach. Irrespective of the optimization technique, the cosine distances consistently remain lower than those from the WNN method. It can also be noticed that LSQP achieved slightly lower similarity than the SQP and hybrid GASQP due to its smaller search space. According to the expert, all 36 counterfactual explanations were coherent with real non-anomalous observations, have similar characteristics as the corresponding anomalous observation and all have been verified to be classified as non-anomalous by the OCSVM model, hence, plausibility and validity criteria are fulfilled. To present the 5-50-95th percentile range improved the interpretability by providing frame of reference. Furthermore, the weighted PCA approach successfully magnifies similarity to the anomaly of regions far from the mean, revealing a weakness in the anomaly detection model, see the plot in the top right corner in Fig. 3, where a pronounced torque peak persists in the counterfactuals at the angular position 1000-1200.

TABLE I. The required execution time to generate the counterfactuals explanations.

Counterfactual explanation approach	Median execution time/ standard deviation (STD) [s]
Hybrid GASQP	16.2 / 1.8
SQP	10.8 / 1.2
L.SQP	4.2 / 0.5
WNN	0.1 / 0.1

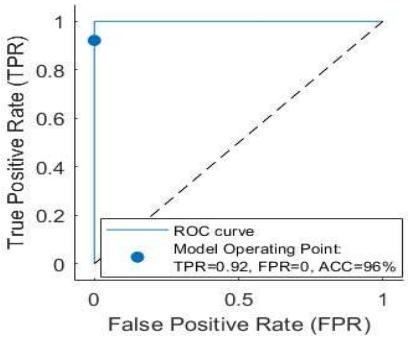

Fig. 2. Results from the OCSVM evaluation.

V. Conclusions

The future goal is to replace human operators in the task of quality control of door locks using a machine learning model that analysis the needed torque to turn a lock 360 degrees. But, to gain the operators' trust in the machine learning model, we need to provide explanations for anomaly classifications. In this study, we evaluated four instance-based approaches for generating human-friendly counterfactual explanations to make the model interpretable. The first approach serves as our baseline, it simply retrieves the nearest non anomalous neighbor in the training data. The other three performs additional feature permutations on the retrieved instance, employing different optimization techniques using a combination of a genetic algorithm and sequential quadratic programming (SQP), SQP alone, and a line bounded SQP, combined with weighted principal component analysis (PCA) to reduce dimensionality and in return the execution time, but also to make the counterfactuals coherent with the real observations. We used the cosine distance between the anomaly and counterfactual explanation to be minimized during the optimization process to achieve similarity and also for the final comparison between the methods. The counterfactual explanations are presented together with the median and 5-95th percentile range to provide a frame of reference of the counterfactuals. Additionally, an expert in interpreting torque data from locks was used to evaluate the plausibility and similarity to real observations. All three-optimization enhanced instance-based approaches successfully created realistic counterfactual explanations, using PCA and cosine distance, with similar results, however line bounded SQP achieved lower similarity. The presentation of the median and 5-95th percentile range further increased the contrast with the anomalous observation compared to using the counterfactual explanation alone, effectively enhancing interpretability of the model and the ability to evaluate the plausibility of the counterfactual explanation. The main difference between the methods is the execution time where line bounded SQP is 61-74% faster (4.2s) than the other two approaches (16.2s and 10.8s) due to the smaller search space. Based on these results it can be concluded that using PCA and cosine distance to measure similarity combined with any of the three optimization approaches described can be a suitable approach to generate counterfactual explanations for an anomaly detection algorithm used with torque data or similar time series data. However, line bounded SQP or SQP bounded by a hyper cube, are preferable in this case due to the faster execution time and simplicity. Furthermore, by adjusting the weights utilized in PCA, a diverse dataset of realistic anomalous observations can be generated. These observations hold potential for enhancing the training and

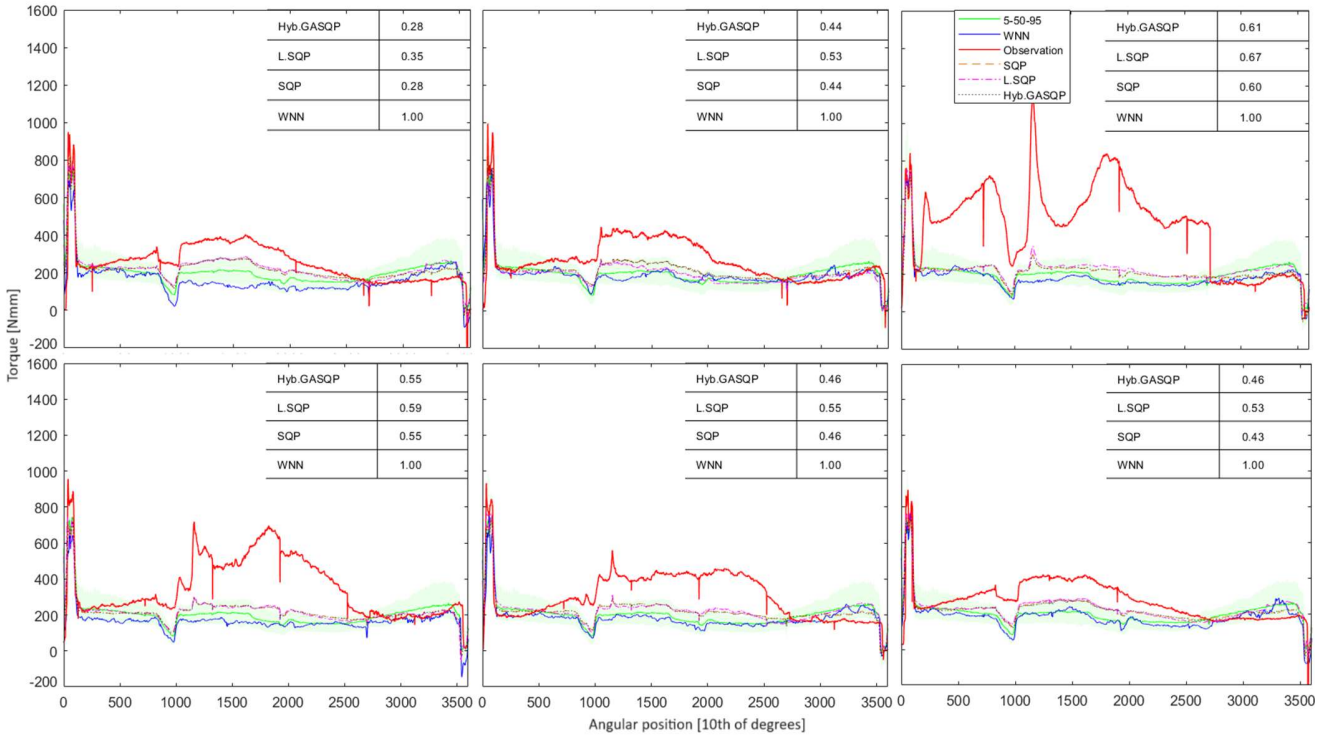

Fig. 3. Here are 6, out of 36, counterfactual explanations and their corresponding anomaly together with the median observation of the valid training data and the 5-50-95th percentile range (shaded region). The relative cosine distance for each approach with respect to distance between the anomaly and the WNN approach's counterfactual explanation can be seen in the corner of each plot.

evaluation of supervised ML models. The incorporation of such counterfactual explanators also imparts valuable insights into the vicinity of the model's decision boundary. This, in turn, renders it a valuable tool for assessing the model's tolerance for unfavorable inputs, potentially revealing model vulnerabilities prior to deployment, as in this study, where a weakness in the anomaly detection model got detected.

VI. LIMITATIONS AND FUTURE WORK

Subjective evaluation of counterfactual explanations by an expert introduces potential bias and risks to validity. The limited sample size of 36 counterfactual explanations may not fully assess the reliability and effectiveness of the used approaches. Exclusively testing on torque data from locks may also limit the external validity of the findings. Evaluating the methods on multiple datasets from different domains would provide a more comprehensive understanding of the performance. Lastly, the use of PCA for dimensionality reduction has resulted in information loss and therefore affecting the sparsity of the results. Further research addressing these limitations could enhance the validity and applicability of the proposed methods in real-world scenarios. Despite these limitations, this study contributes to interpretable machine learning for anomaly detection in industrial settings.

ACKNOWLEDGMENTS

This project has received funding from The Knowledge Foundation, Mälardalen University and Assa Abloy under grant agreement No 20200132 01 H.

REFERENCES

[1] S. Russel and P. Norvig, Artificial Intelligence A Modern Approach, 3rd ed. New Jersey: Prentice Hall, 2010.

[2] A. B. Nassif, M. A. Talib, Q. Nasir, and F. M. Dakalbab, "Machine learning for anomaly detection: A systematic review," IEEE Access, vol. 9, pp. 78658–78700, 2021.

[3] N. Seliya, A. Abdollah Zadeh, and T. M. Khoshgoftaar, "A literature review on one-class classification and its potential applications in big data," J Big Data, vol. 8, no. 1, pp. 1–31, Dec. 2021.

[4] Feng Pan, Wei Wang, A. K. H. Tung, and Jiong Yang, "Finding representative set from massive data," in Fifth IEEE International Conference on Data Mining (ICDM'05), IEEE.

[5] A. Sgueglia, A. Di Sorbo, C. A. Visaggio, and G. Canfora, "A systematic literature review of IoT time series anomaly detection solutions," Future Generation Computer Systems, vol. 134, Sep. 2022.

[6] C. Molnar, Interpretable Machine Learning, 2nd ed. Independently published, 2022.

[7] P. Lipton, "Contrastive explanation," Royal Institute of Philosophy Supplement, vol. 27, pp. 247–266, Mar. 1990.

[8] R. Guidotti, "Counterfactual explanations and how to find them: literature review and benchmarking," Data Mining and Knowledge Discovery, Apr. 2022.

[9] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," SSRN Electronic Journal, 2017.

[10] T. Andersson, M. Bohlin, M. Ahlskog, and T. Olsson, "Sample size prediction for anomaly detection in locks," in 56th CIRP Conference on Manufacturing Systems, CIRP CMS, Cape Town: Elsevier, Oct. 2023.

[11] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in AI," in Proceedings of the Conference on Fairness, Accountability, and Transparency, in FAT* '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 279–288.