How to Explain and Justify Almost Any Decision: Potential Pitfalls for Accountability in AI Decision-Making

Joyce Zhou
Cornell University
Ithaca, NY, USA
jz549@cornell.edu

Thorsten Joachims
Cornell University
Ithaca, NY, USA
tj36@cornell.edu

ABSTRACT

Discussion of the "right to an explanation" has been increasingly relevant because of its potential utility for auditing automated decision systems, as well as for making objections to such decisions. However, most existing work on explanations focuses on collaborative environments, where designers are motivated to implement good-faith explanations that reveal potential weaknesses of a decision system. This motivation may not hold in an auditing environment. Thus, we ask: how much could explanations be used maliciously to defend a decision system? In this paper, we demonstrate how a black-box explanation system developed to defend a black-box decision system could manipulate decision recipients or auditors into accepting an intentionally discriminatory decision model. In a case-by-case scenario where decision recipients are unable to share their cases and explanations, we find that most individual decision recipients could receive a verifiable justification, even if the decision system is intentionally discriminatory. In a system-wide scenario where every decision is shared, we find that while justifications frequently contradict each other, there is no intuitive threshold to determine if these contradictions are because of malicious justifications or because of simplicity requirements of these justifications conflicting with model behavior. We end with discussion of how system-wide metrics may be more useful than explanation systems for evaluating overall decision fairness, while explanations could be useful outside of fairness auditing.

CCS CONCEPTS

• **Social and professional topics** → **Computing / technology policy**; • **Human-centered computing** → **Interactive systems and tools**; • **Information systems** → *Decision support systems.*

KEYWORDS

explainable AI, right to an explanation, adversarial explanations

ACM Reference Format:
Joyce Zhou and Thorsten Joachims. 2023. How to Explain and Justify Almost Any Decision: Potential Pitfalls for Accountability in AI Decision-Making. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3593013.3593972

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06…$15.00
https://doi.org/10.1145/3593013.3593972

1 INTRODUCTION

There has been growing discussion of the "right to an explanation" for people subject to partial or fully automated decisions. This includes but is not limited to clear references in the European GDPR, the proposed Canadian privacy bill C-11, and in calls for research and discussion in this topic [15, 35, 38]. However, these legal bills do not clarify what goals such an explanation should serve to fulfill, what an "explanation" precisely is, or how to reliably distinguish an "explanation" from a "justification" or "rationalization". What should an explanation that is created to fulfill this "right to an explanation" aim to communicate, what standards should we have for this kind of explanation system, and how do we judge whether this "right" has been adequately met? Finally, how would fulfilling this "right to an explanation" to those affected by an automated decision benefit them or address the problems that are motivating these discussions? How does such an explanation fit into a greater decision system auditing environment?

Within computer science, we are only starting to understand how to build trustworthy AI systems. Explanations are seen both as a way of holding decision makers accountable, and as a way of generating trust among those on the receiving end of these decisions. However, while AI experts may use explanations frequently to probe the inner workings of the systems they build, decision recipients are less likely to be familiar with the AI system details or to have general knowledge around the decisions themselves. Simple explanations may offer, for example, a criminal defendant some insights into how their risk assessment score was computed or how it fits into other sentencing guideline systems. But the precise nature of such explanations raises complex questions about the requirements, goals, and standards.

In this complex space of questions, we provide a negative result where a seemingly well-designed and externally verifiable explanation system can hide the true nature of the underlying decisions. We show that simply maintaining a "right to an explanation" is not enough to identify malicious decision systems, even if we impose specific verifiable quality requirements on the explanations. In particular, we require that the explanations are sufficiently simple to be understandable and that they fulfill statistical-significance requirements on public data. By defining and building an example explanation system meeting these criteria and applying it to simplified data centered around COMPAS recidivism prediction [18], we demonstrate how such an explanation system could be abused to defend even a severely unfair decision system[1]. This echoes existing issues that have been raised about explainability and black-box

[1]Source code is publicly available at https://doi.org/10.5281/zenodo.7192644.

models [33], and further highlights the gap between servicing decision recipients on an individual level and system-wide decision auditing.

Specifically, we focus on a situation in which the group making a decision using the outputs from a recidivism prediction model is also responsible for providing an explanation. As such, they are motivated to present explanations that defend whatever decisions they make. For clarity, we refer to this as a "*justification*" instead of an "*explanation*" to emphasize its purpose in defending a decision made, in contrast to providing a good-faith visualization of the decision system itself [2]. We treat both the decision model and the justification system as black boxes, where outsiders only inspect the decisions and justifications provided, and not how the systems themselves function internally.

We find that regardless of how relevant the features referenced by a justification system are to the underlying model, or how accurate that model is, the majority of decisions could be defended with a justification that appears statistically significant and supports whatever decision was made. At a simplified case-by-case level, most decisions could be defended by some kind of justification, and a critical 68% majority of cases could be justified for any decision the system may chose to make. If we audit the justification system itself for its faithfulness to the decision system based on the decisions and justifications made across multiple cases, conflicts between provided justifications become more visible. However, it is hard to tell whether these conflicts exist because the justifications are maliciously defending a discriminatory model, or whether they are made in good faith but still differ from the original model because they are simplified for readability. There appears to be no intuitive faithfulness threshold that reveals whether a justification system is covering up any intentionally discriminatory decision system. We argue that in a real-world situation, it may be more effective to audit decision systems by comparing the effectiveness of different potential justifications proposed by auditors themselves using publicly accessible decision contexts and outcomes, instead of prioritizing justifications provided by a decision-maker. While individual decision recipients may still be able to benefit from explanations given in good faith, it is worth remembering that not all decision makers work in good faith and a a malicious system requires more than individual responses to recognize.

2 RELATED WORK

So far, most work on AI interpretability or explainability has centered around people who are developing AI systems and interfaces. For instance, explanations may be designed for AI experts and data experts who may be structuring and evaluating the model itself [27, 30], or AI novices who are end users of such systems being given assistance through AI decision-making [20, 22, 30, 39]. Explanations presented to experts designing and debugging a system might be evaluated by how well they expose biases within the system [1, 10, 30], what types of input flaws may be revealed [1, 32], or how they handle edge or adversarial cases [1, 14, 21, 25]. For explanations presented to assist human-AI team decision-making, evaluations are frequently centered around appropriate trust [3, 16, 30], mental models [23, 30], or overall team performance [3, 30].

As part of these goals, metrics and higher-level goals for explanation quality that seem to support improved model property discovery or decision team experience have been suggested. These include but are not limited to simple and understandable explanations [29, 31], soundness, completeness, or faithfulness of explanations [19, 40], or formalizing interpretability itself and presenting methods to evaluate it across varying model classes and tasks [11]. There has also been some focus on how the presentation of the decision model and explanations [7], or their relationships to the task [17] affect human decision-making and overall trust. Finally, there have been efforts to design quantitative metrics for how faithfully explanations reflect underlying model behavior [5, 9, 36].

Outside of the model creation and usage process themselves, explanations have been suggested and critiqued as potential tools for auditing model performance and final decisions [6]. For example, [14] discuss how saliency maps (highlighting important areas of an image) are commonly used as explanations with medical image analysis systems, but fail to help with adversarial input analysis and could be misused to make a model seem more or less effective than it really is. In general, the legal right to an explanation has been highlighted as a way to identify unethical or unacceptable AI systems [15] and providing some base to make decision objections off of [35, 38]. However, it is still unclear what requirements an explanation satisfying this right would have to satisfy [12, 13], or if providing an explanation would help these goals at all [13].

Here, we focus on the concern that manipulative explanation systems could intentionally support or defend a system. We know a human decision maker could make a decision first and come up with some way to rationalize their decision afterwards that is hard to prove anything about. What is preventing AI systems from doing something similar, and how could we detect if they are [24]? There is some existing work that highlights this same issue. [37] presents an adversarial model that could fool LIME input perturbation sampling in order to present explanations for racist decisions that focus on innocuous features. Similarly, [2] demonstrates how unfair models could be presented with maliciously generated fair rule lists that still appear faithful to the model itself. Here, we focus on how simple explanations could fail to identify racist models in auditing scenarios where the burden of explanation is on the decision-maker despite them seemingly matching past data, and how individual decision recipients vs. a larger audience are able to respond.

It is still important to keep in mind that explanations are not the only tools available for auditing models, especially if we collect a larger number of decision outcomes. Fairness goals such as demographic parity or predictive equality are evaluated based on a larger set of model decisions and outcomes [8, 28]. Auditors can also examine for biases in the data used for model training [28], although this is more difficult if the actual dataset is private and all we have access to is public information.

[2]Other works often use "explanation" in a way that includes this kind of black-box explanation and decision system.

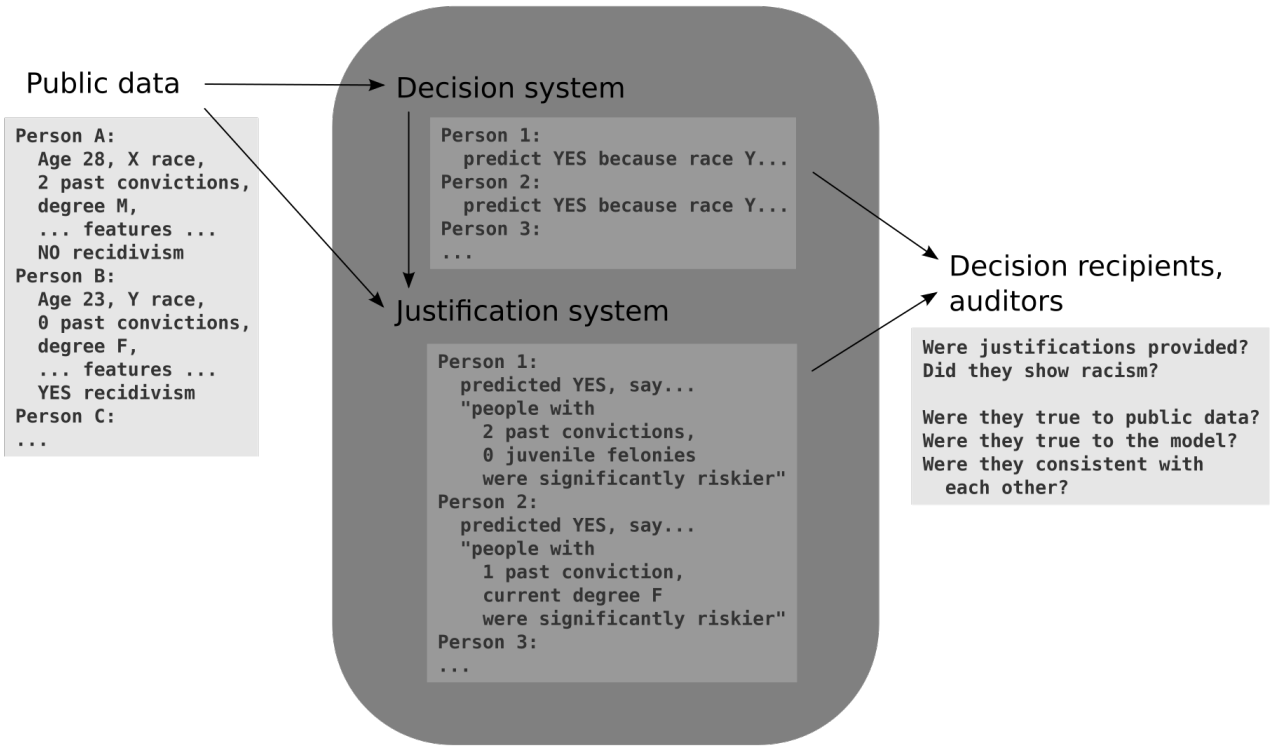

Figure 1: We examine scenarios where the decision-maker is responsible for fulfilling the "right to an explanation". The area with a dark grey background represents the black-box systems that are privately designed, implemented, and owned by the decision-maker. These black-box systems are designed with the intention of making decisions and defending the decision-making model as ethical. The area with a light grey background represents what information is publicly accessible, as well as examples of public auditing questions that might be asked. The public information and audit questions are made with the motivation of revealing any unethical model decision processes.

3 EXPERIMENT: HOW CAN JUSTIFICATIONS BE MANIPULATED WHEN WE ONLY EXAMINE THEM CASE-BY-CASE?

We first examine the potential for justifications to defend decisions on an individual case-by-case basis, to simulate the scenario where those being given decisions and justifications can't or don't communicate with each other. For example, they may not have access to contact information for similar decision recipients, so they never think to reach out to others. Privacy concerns can convince recipients to avoid disclosing decision information, justifications, or potentially relevant personal context for these decisions. Even if information is shared somewhere, it may not be collected easily.

This is an intentionally simplified scenario, as a real-world decision system with high demand for justifications would likely have some related discussion community where decision and justification information is shared. However, we start with this question to examine simple weaknesses of case-by-case justifications, as well as to set up the framework of a manipulative justification system.

How many individual decision recipients could be given simple, verifiable, and relevant justifications that defend whatever decision they were given? In Figure 1, we show the overall structure of this scenario. Areas with light background are visible to decision recipients and auditors, while areas with dark grey background are the black-box systems maintained by the decision-making group with motivations to defend their systems.

To do this, we started by outlining the requirements for simple, verifiable, and relevant justifications for a given decision.

3.1 Justification Requirements

Decision auditors and decision recipients benefit from receiving usable justifications, so we define a set of requirements: they need to be simple, they need to be verifiable, and they can only be applied to a decision if they are relevant. We assume that if a justification meets these requirements, then it appears satisfactory to a decision recipient.

For a justification to be **simple**, it needs to be easily understandable. If an explanation or justification is not simple enough to understand, then it is effectively useless [29, 31]. There is not a common threshold for adequate simplicity, so we call a justification simple if the number of features it references is under a fixed threshold. That is, for some low constant N_f, the number of features mentioned in a justification c satisfies the condition $c \leq N_f$.

For a justification to be **verifiable**, it needs to appear true to a decision recipient. As most past work focuses on explanations for

developers or collaborators, this is typically not a clear priority or is assumed to hold because the explanation system has direct access to the model [26, 32]. In our scenario, we call a justification verifiable if it can be confirmed based on past public records. Because we assume that decision recipients are not always able to access information about other decision recipients, this is the most relevant information available to them.

For a justification to be **relevant**, it can only be used for a decision if their conditions match up. For example, if a justification captures features that do not exist for a decision, or if it supports a different final outcome, then it is not relevant.

The decision-maker benefits from visibly meeting these requirements but still having flexibility to work around them, so we outline a simple but flexible justification template. We used justifications that contain up to some number N_f of features and identify if all people with the same values for those features were significantly[3] more or less likely to recidivate compared to the entire arrest population in this dataset. For example, one justification might be that "people currently charged with a felony and with 2 prior juvenile felony convictions are significantly less likely to recidivate compared to the overall group". We assume a decision recipient or system auditor knows little about how the decision or justification system works internally, so they would be unable to directly compare these dataset features with the system configuration and are only able to verify the external criteria.

Justifications using this template are simple because they contain a limited number of features. They are verifiable because they can be confirmed using a statistical significance test on a past dataset. Finally, they are relevant when their feature values and final significance comparison match up with the features and decision model output of a specific case.

This justification template is dependent on a dataset of past recidivism outcomes for verifiability, as well as a set of "current" model decisions that need to be justified. We use the COMPAS recidivism dataset to serve as both a "reference" for justification verifiability, and the source of test cases for how well justifications can defend a range of decision models.

3.2 Dataset

To build the dataset supporting model decisions and potentially usable justifications, we used the ProPublica COMPAS two-year-recidivism dataset [18]. We focus on recidivism prediction specifically because it greatly impacts the lives of decision recipients, while the prediction systems (such as COMPAS) are often privately maintained and not well-understood by decision recipients.

This dataset contains information about 7214 people who were assessed using COMPAS scores in the pretrial process of criminal defendants in Broward County, Florida, USA, from 2013 to 2014. It contains personal information (names, birth date, legal sex, race), criminal records information (age at arrest, number of prior misdemeanors and/or felonies, relevant criminal charge at pretrial time), COMPAS scoring information (COMPAS decile score, simplified high- vs low- risk recidivism prediction), and 2 year recidivism outcome information. Similar to ProPublica, we filter the dataset to exclude cases that were ordinary traffic offenses, missing COMPAS decile scores, or had charge and arrest dates over 30 days apart. This leaves us with a dataset containing case information and recidivism predictions for 6172 people.

Finally, we did an 80/20 split of the filtered dataset into reference and test sets, leading to a reference population with size 4937 and test population with size 1235. The reference set is used to identify which potential justifications are verifiable. The test set is used to simulate decision models and evaluate how well a justification system would defend them.

Next, we describe how we used the reference dataset to generate all potentially usable (simple and verifiable) justifications.

3.3 Usable Justification Generation

To emulate the decision-maker, we abuse the multiple comparisons problem to identify every usable justification (following the template) for each decision. One of these is presented to the decision recipient as a "final" justification, with its verifiability based on one statistical significance test.

To generate a list of all possible justifications across all cases, we iterated through every possible combination of feature values for up to N_f usable features, identified the subset of reference population cases that matched those features, and calculated whether those reference population cases had actual recidivism rates significantly higher than the overall reference populations without any multiple comparisons correction. A justification was deemed significant if and only if the Clopper-Pearson confidence intervals (using $\alpha = 0.05$) for subset and general recidivism rates did not overlap. By default, we considered only the "juvenile felony count", "juvenile misdemeanor count", "juvenile other conviction count", "total prior conviction count", "charge degree", and "charge description" fields as usable for justification[4].

3.4 Decision Systems

How well can this manipulative justification system defend extremely biased or random decisions? We simulated four risk assessment decision systems to test this justification system on. Each decision system classifies a case as either "low-risk" or "high-risk".

- The **Original** decision system uses the low-risk and high-risk recidivism predictions from the simplified ProPublica COMPAS dataset.
- The **Racist** decision system sorts cases by defendant race and classifies them as low-risk and high-risk based on that ranking, with the same percentage of low- vs. high-risk decisions as the original system.
- The **Oracle** decision system has perfect accuracy, classifying defendants as low-risk and high-risk based on whether they actually recidivated within two years of arrest.
- The **Random** decision system randomly classifies defendants as low-risk and high-risk, with the same percentage of low- vs. high-risk decisions as the original system.

[3] Without multiple comparisons correction, as the decision recipient receiving this justification is unsure how many significance tests may have been done when calculating a justification.

[4] Using the protected traits ("age", "age category", "sex", or "race") as part of a decision justification would be visibly unethical or illegal compared to justifications that only use non-protected traits.

Table 1: Justifiability of model decisions, varying across decision models and justification complexity.

N_f	Model	None	Both	Support	Against	% Justif.
$N_f \leq 2$	Original	0	846	284	105	91.49%
	Racist	0	846	243	146	88.17%
	Oracle	0	846	286	103	91.65%
	Random	0	846	182	207	83.23%
$N_f \leq 3$	Original	0	850	281	104	91.57%
	Racist	0	850	241	144	88.34%
	Oracle	0	850	283	102	91.74%
	Random	0	850	181	204	83.48%

3.5 Justifiability Metrics

Finally, we measured how many cases could be defended by counting the percentage of cases that have any usable justification. Because we assume each decision and justification is being examined at a case-by-case level, it does not matter which exact justification is being presented for each case: as long as there is at least one, the final decision could be defended somehow.

If a case only has usable justifications that agree with a model decision, it is deemed "justifiable" at the case-by-case level. If it only has usable justifications against the model decision, then it is not. If it has usable justifications available for both "low-risk" and "high-risk" decisions, then it also counts as "justifiable". These cases are especially interesting: a case with usable justifications for any possible decision can be defended at a case-by-case level regardless of the actual decision.

3.6 Results: Case-by-case

We now look at how successful this manipulative justification system is when attempting to defend decisions made by the original, racist, oracle, and random decision systems on the test set cases.

In Table 1, we count how many cases in the test set have justifications that could defend any model decision, compared to having only justifications in favor of or against the actual predictions from each model. In the overall test population, over 90% of original model decision cases have some justification usable in favor of the actual decision. The racist and random models had slightly lower justifiability percentages, although all of these model predictions were still over 80% justifiable. However, notably, *over 68% of all test set cases have applicable justifications that could be used to defend both the "low-risk" and "high-risk" decisions, which means any decision model is justifiable on these cases*! The differences in justifiability between models is solely caused by the set of cases where justifications are only available in favor of one potential decision, and what decisions are given by a particular model on those cases.

If we increase how many features a justification is allowed to use (and how complex a justification is allowed to be in general), both these percentages increase very slightly. This can be explained by a combination of more feature sets available to use for justification, but fewer of those explanations reaching the significance threshold after splitting the data even more finely. As a result, including even more complex explanations is unlikely to provide additional expressive power.

Again, we emphasize that based on Table 1, *there is at least a 68% majority of cases that can always be defended in a case-by-case scenario regardless of the decision model because they have justifications available for any decision.*

So when decision recipients are unable to communicate decision and justification details with each other, the majority of them could be given justifications that defend a decision, regardless of how accurate or fair the decision system itself is. However, for real-world decision systems, this is usually not the case. Decision recipients may well be able to reach out to each other and form communities. Furthermore, they benefit from discussing and revealing whenever a decision or justification system is being manipulative. Auditing across multiple decisions and justifications with public records data is something that would be a simple first step towards defending them. So how well might this work?

4 EXPERIMENT: COULD CHECKING FOR SYSTEM-WIDE JUSTIFICATION FAITHFULNESS HELP IDENTIFY MANIPULATIVE SYSTEMS?

If we have access to multiple decision recipients' decision and justification information, we could evaluate justifications using metrics for overall justification system faithfulness. We now simulate a scenario where justifications for each decision are independently provided, but all case data, decisions, and justifications are publicly visible for auditing. To do this, we emulate the decision maker by running a justification-providing system that assigns each test set decision case a justification independently of every other case, and emulate the decision auditors by calculating global consistency, global sufficiency, and uniqueness metrics from [9]. Given the justifications that a justification system gives in defense of some decision system, could we assess how faithful or manipulative the justifications are, and would this identify malicious decisions or justifications?

4.1 Faithfulness Metrics

To emulate decision auditors and decision recipients, we used faithfulness metrics proposed by [9] as a way to measure how internally coherent and reasonable a justification system seems, based on what justifications it provides for a set of decision cases. It features two metrics (consistency, sufficiency) that measure whether or not provided justifications can contradict with each other based on what outcomes the relevant decision cases got, as well as a uniqueness metric that measures how many repeated patterns there are across the justifications provided. The malicious justification system aims to have high consistency and sufficiency and low uniqueness in order to appear reasonable to decision auditors, while using only acceptable features to justify as many cases as possible.

For a justification system to have high **consistency**, cases that are assigned the same justification should have similar outcomes. It can roughly be summarized as "the expected fraction of cases given the same justification, across the justification for each case, that got the same decision outcome". If a system has low consistency,

it implies that the same justifications are being used for decisions that frequently contradict each other.

For a justification system to have high **sufficiency**, cases that are relevant to the same justification (even if they were not assigned that justification) should have similar outcomes. It can roughly be summarized as "the expected fraction of cases that the same justification is applicable to, across the justification for each case, that got the same decision outcome". If a system has low sufficiency, it implies that there are justifications that could cover cases with decisions that contradict each other, even if they are never officially applied.

For a system to have low **uniqueness**, there should be few cases assigned justifications that are never used elsewhere. Uniqueness is calculated as the fraction of decision cases assigned a justification that was assigned to no other observed case. If a system has high uniqueness, it means that a large number of decisions are justified with something that is never repeated elsewhere. In the worst case, if a system has 100% uniqueness, then every justification is unique: even if these justifications are technically true, they are extremely unhelpful for identifying common patterns across different cases.

We now describe how these metrics are calculated. Global consistency is defined as:

$$m^c = \mathop{\mathbb{E}}_{x \in \mu\mathcal{X}} \left[\Pr_{x' \in \mu C_{\pi=e(x)}} \big(f(x') = f(x)\big) \right]$$

Global sufficiency is defined as:

$$m^s = \mathop{\mathbb{E}}_{x \in \mu\mathcal{X}} \left[\Pr_{x' \in \mu F_{\pi=e(x)}} \big(f(x') = f(x)\big) \right]$$

Global uniqueness is defined as:

$$m^u = \frac{|\{\, x \in \mathcal{X} : |C_{e(x)}| = 1 \,\}|}{|\mathcal{X}|}$$

where:

- $\mathcal{X}$ is the full set of decision cases in the test set.
- $f(x)$ is the decision made for case x.
- $e(x)$ is the justification selected for case x.
- $C_\pi = \{x \in \mathcal{X} : e(x) = \pi\}$ is the set of all cases that the justification π was assigned to.
- $A(x, \pi)$ is true if and only if justification π could describe case x, even if its claim differs from the decision made.
- $F_\pi = \{x \in \mathcal{X} : A(x', e(x))\}$ is the set of all cases that the justification π could describe.
- μ is a probability distribution, which we treated as uniformly distributed when calculating these metrics.

Note that these metrics evaluate a justification system that gives one justification for each decision case, and can vary depending on the exact cases and justifications that are given.

In the case-by-case scenario, we assumed that if there is any usable justification, then a decision case is justifiable because there is no inter-case communication. However, in a system-wide faithfulness check, this no longer holds and we need to implement some justification system that selects exactly one justification to give for each decision case.

4.2 Justification System

To emulate the decision-maker, we outline a justification system that meets simplicity requirements and aims to maximize faithfulness metrics, while still maintaining justification flexibility. We assume its developers are not aware in advance exactly what cases and decisions they will need to generate justifications for, but they are aware of what metrics will be used. All they have access to is a representative (reference) set of past cases, the decision model itself, and decision model predictions for both past cases and the current case they need to provide a justification for. They want to present the decision model as a reasonable model that does not use protected features.

In this experiment, the justification system selects one usable justification to give for each decision case in the test set. For each case, there is a set of usable justifications with feature values that match up that could be used for that case (not necessarily matching on decisions, as there are some cases that only have usable justifications in favor of one decision). Likewise, for each usable justification, there is a set of cases that has matching features (again, not necessarily matching decisions). Thus, for each decision case, we must select one of the usable justifications as the "final" justification. Because we assume the justification system designers are aware what metrics they are audited with, we select a relevant justification independently for each decision case that naively maximizes on faithfulness metrics. Ideally, the justifications selected defend the decision model as much as possible.

We implemented a ranking system that selects a justification that primarily defends the decision that was made, and otherwise prioritizes minimal disagreement between predictions based on estimates from the relevant reference set cases. For some of the test cases, there were usable justifications only available in defense of one potential decision. If there is no usable justification that defends the relevant case decision, it either gives an opposing justification with the fewest decision conflicts with the idea that some kind of justification is mandatory (a "must-justify" system), or no justification at all with the premise that "there is no simple way to defend this decision" (an "agree-only" system).

For the "must-justify" justification system variant that occasionally gives opposing justifications, we included all cases in the metrics. For the "agree-only" justification system variant that occasionally fails to give any justification at all, we excluded those cases from the metrics. Thus, we also show the "% Justified" metric for how many cases received a justification with this system at all. Note that because an "agree-only" justification selection system would provide a justification if and only if there is one that would support the decision made, all justifications are only used in favor of decisions they agree with. Thus, the consistency metric for a "agree-only" justification system always equals 1, regardless of what decisions it is defending.

4.3 Results: System-wide faithfulness

We ran both variants of the justification system together with all decision model variants on the test set, and calculated faithfulness metrics for each combination.

In Table 2, we compare faithfulness metrics based on what justifications would be given by the "agree-only" justification selection

Table 2: System-wide faithfulness using only "agree-only" justifications, varying across decision models.

N_f	Model	% Justif. (↑)	Cons. (↑)	Suff. (↑)	Uniq. (↓)
$N_f \leq 2$	Original	91.49%	1.0000	0.6674	0.0221
	Racist	88.17%	1.0000	0.6117	0.0257
	Oracle	91.65%	1.0000	0.6646	0.0203
	Random	83.23%	1.0000	0.4929	0.0311
$N_f \leq 3$	Original	91.57%	1.0000	0.6707	0.0415
	Racist	88.34%	1.0000	0.6174	0.0476
	Oracle	91.74%	1.0000	0.6682	0.0388
	Random	83.48%	1.0000	0.4980	0.0514

Table 3: System-wide faithfulness using only "must-justify" justifications, varying across decision models

N_f	Model	% Justif. (↑)	Cons. (↑)	Suff. (↑)	Uniq. (↓)
$N_f \leq 2$	Original	100.00%	0.9049	0.6481	0.0210
	Racist	100.00%	0.8935	0.5954	0.0226
	Oracle	100.00%	0.9067	0.6446	0.0194
	Random	100.00%	0.8549	0.4934	0.0259
$N_f \leq 3$	Original	100.00%	0.9890	0.6512	0.0388
	Racist	100.00%	0.9889	0.6006	0.0421
	Oracle	100.00%	0.9875	0.6480	0.0364
	Random	100.00%	0.9858	0.4976	0.0429

system in defense of the original, racist, oracle, and random decision models. We can see that the sufficiency metric for justifications across all four decision systems is startlingly low. An intuitive interpretation of the sufficiency metric for the original decision model is "the average fraction of cases that each justification could apply to and would agree with the final decision of was only 67%". In other words, most of the time, the justifications that were given could easily apply to other cases that had different outcomes. However, this is also a side effect of using simple, relatively interpretable justifications. If we allow more complex justifications, then the justifiable case fraction and sufficiency improve while uniqueness worsens. This pattern holds across multiple decision systems.

A similar pattern happens in faithfulness metrics for a "must-justify" justification system, except in these the "% Justified" metric remains fixed at 100% and consistency varies instead. The same low overall sufficiency and trade-off between uniqueness and other metrics remain. We show these results in Table 3.

Similar to before, these faithfulness metrics vary across decision models and can indicate how this justification system matches better with the original decisions than racist or random decisions. However, it is also unclear what a reasonable threshold for consistency, sufficiency, or uniqueness may be. Justifications for all but the random decision system show high consistency, low uniqueness, and sufficiency above 0.5.

If it is difficult to determine an intuitive threshold we can use for justification faithfulness but we still have access to a public case dataset in this system-wide scenario, how else could decision auditors use justifications to identify malicious decision systems?

Auditors could try to determine the validity of the justification itself by comparing its faithfulness across different potential decision systems, as presented in these results tables. However, contrasting these requires that auditors can somehow recreate justifications for any potential decision input and outcome, not just the ones that already exist. Furthermore, this ultimately serves to understand the justification system itself better, and not the original decision model we want to audit.

Instead, auditors could contrast the faithfulness of different potential justification systems that they build themselves, against a given decision system and its past outputs. One challenge that comes up with this approach is that we encounter a trade-off between uniqueness vs. consistency and sufficiency: it is hard to control for justification system uniqueness when we allow justifications with varying structures or features. To demonstrate potential benefits and downsides of this contrast approach, we run the same justification system but with the additional "race" feature allowed in a justification template. This represents a situation where the decision auditors build their own justification system to challenge the officially provided justifications, in order to test which one proves more faithful.

We contrast faithfulness metrics from this extended justification with those of the original in Table 4. We see a large gap between race-using and race-excluding justification faithfulness for the "racist" decision model, with sufficiency especially increasing sharply. However, justifiability, uniqueness, and sufficiency all increase slightly across all of these decision model contrasts. While the difference is sharpest for the "racist" decision model, simply identifying an increase in justifiability, consistency, or sufficiency is still ambiguous. Furthermore, increases in consistency and sufficiency seem to correlate with increases in uniqueness as well - this is the trade-off challenge mentioned earlier.

Thus, while contrasting different potential justifications on faithfulness metrics may help identify flaws in these systems, it is still unclear how to handle the consistency/sufficiency and uniqueness trade-off, as well as what causes these changes in metrics. Furthermore, these contrasts do not need to use any official justification source. In fact, such a comparison could be done without any "right to explanation" at all: as long as there is a collection of decision cases and their outputs, auditors could hypothesize a range of justifications and evaluate them.

5 DISCUSSION

In the vast majority of test cases in this recidivism prediction dataset, it is possible to provide a bad-faith justification at a case-by-case level that still appears simple and verifiable by taking advantage of the multiple comparisons problem. For a critical 68% majority of cases, it is possible to do this for *both* potential predictions: regardless of what prediction a malicious decision-maker is defending, there is a cherry-picked statistical comparison available for them to use. On a more complex dataset (e.g. with more fields that may not be directly interpretable), we speculate that the percentage of justifiable cases would only increase.

Overall, the right to an explanation could easily be abused to defend decision models in standalone cases if we have no clear definition of what an explanation should address or a clear way to

Table 4: System-wide faithfulness metrics if we include "race" in a justification vs. not (using only $N_f \leq 3$ "agree-only" justifications)

Model	Use Race?	% Justified (↑)	Consistency (↑)	Sufficiency (↑)	Uniqueness (↓)
Original	No	91.57%	1.0000	0.6707	0.0415
	Yes	95.30%	1.0000	0.6868	0.0756
Racist	No	88.34%	1.0000	0.6174	0.0476
	Yes	99.43%	1.0000	0.9227	0.0643
Oracle	No	91.74%	1.0000	0.6682	0.0388
	Yes	95.30%	1.0000	0.6811	0.0747
Random	No	83.48%	1.0000	0.4980	0.0514
	Yes	89.39%	1.0000	0.5025	0.0860

audit the explanation generation process. The justification template we used is based on data in the same distribution that the decision model was trained on, and is arguably still connected to the model itself, but fails to accurately represent the model internals or answer questions like "what factors caused the model to predict X instead of Y?". Instead, it presents something like "prediction X from the model may be reasonable because of these factors". Developing clearly defined standards for explanation complexity [29, 31], soundness and completeness [19], creation process and burden of responsibility, what data an explanation should have access to, or other auditing mechanisms might help with this. However, there will likely still be unintentional or malicious cases where these standards fail to keep decision systems accountable.

If we assume auditors have access to multiple decision cases and justifications, it becomes harder to attempt justifying an entire group of test cases without creating conflicts between justifications. A justification used in defense of one case may be applicable to and conflict with the prediction of another case, while avoiding this kind of conflict may lead to an increased number of cases without any justification at all. We can capture this conflicting behavior using justification (explanation) faithfulness metrics. However, the complexity of these metrics also makes it hard to tell what a natural threshold for faithfulness is. Is decreased consistency or sufficiency more a result of requiring simple justifications, or is it more a side effect of the justification system being manipulative and hiding the usage of protected features?

In our experiments, the clearest indicator of malicious decision systems came from a contrast between faithfulness metrics of different candidate justification systems. Interestingly, this kind of comparison requires no "right to an explanation" at all - instead, it relies on having an accessible dataset of decision case contexts and outputs. This indicates that the "right to an explanation" alone is not necessarily helpful for verifying the validity of decision systems. If we are checking for overall system validity, it may be more effective to enable full audits from outside observers with accessible decision outputs.

Our findings echo and contribute to the body of existing work, in that they highlight how explanations can be easily manipulated to hide heavy biases in a decision system even if externally verifiable against a public dataset. We do acknowledge that the explanation system and format used here not a state-of-the-art method. However, it is ultimately still decision makers likely creating explanations, and decision recipients who are shown these explanations. If recipients are unfamiliar with the explanation creation process, this difference may not be obvious to them. If an explanation for the recipient includes information about how it was generated [34], this places more burden on the recipient to understand and respond to, as well as leaving a door open for yet another false explanation (justification).

While there may still be ways to make use of explanation systems operated by decision makers in automated decision-making, such as highlighting decision feedback or adjustment mechanics [38], using them for system auditing is not necessarily trustworthy or reliable. Instead, open communication between decision recipients and third-party examination across multiple cases may be more effective for system auditing. Furthermore, instead of prioritizing an explanation of the reasoning going into the construction of a decision-making system (or the construction of the explanations themselves [4]), which gives decision makers the power to defend an existing system, it is worth considering the overall decision outcomes and impacts when auditing.

In the future, it would be interesting to explore different explanation system quality metrics, especially as explanations are still useful outside of auditing. Specifically, we could contrast how explanation systems (malicious or good-faith) may present explanations across multiple decision cases at a more detailed level. For example, we could aggregate multiple explanations presented and analyzing why they agree or conflict on similar inputs, to evaluate the overall usefulness of the explanation system. This kind of contrast has been used as criticism in past work [14], but it is an open question how well it could be used as a general metric for evaluating explanation systems.

6 ETHICS

This work did not rely on any real-life deployed systems, direct human interaction, or private datasets. All experiments we ran were simulations with representations of what motivations each involved party may have, based on publicly accessible criminal records data that has been frequently used in research of decision fairness and legal guidelines.

However, as we demonstrated a naive method for malicious decision-makers to provide malicious justifications in response to the "right to an explanation", this could enable or encourage similar behavior in real-life decision-making and explanation generation. We considered this risk and decided it was worth accepting, in a similar spirit to security vulnerability research. We decided that it was unlikely for any malicious decision-makers to begin providing malicious justifications based on this work without previous intentions to, and that it was very unlikely that the naive malicious justification selection algorithm we present would provide anything useful for such a group.

7 CONCLUSION

We simulated two scenarios based on COMPAS recidivism prediction where the risk assessment decision-maker is obligated to fulfill a "right to an explanation" for their decision recipients, and tries to defend as many of their decisions as possible. As part of this "right to an explanation", decision-makers needed to use "explanations" that appear simple and verifiable to their decision recipients.

In the first scenario, we assumed that decision recipients were unable to communicate with each other. We found that if the decision-maker takes advantage of the multiple comparisons problem, for the majority of decision cases, they are able to provide a malicious justification in the form of "past cases with these small number of matching features were significantly more or less likely to reoffend than the general arrest population". This is true regardless of what decision model is actually being used - a model with perfect accuracy, a model solely based on race, or even a random model all have a majority of justifiable cases.

In the second scenario, we assumed that decision recipients were able and willing to communicate their decision cases, results, and provided justifications with each other. We measured justification quality across multiple cases using faithfulness metrics, and found that they did vary across different justification system and decision system combinations. However, there was not an intuitive threshold to determine whether a justification system is maliciously defending a decision system. Furthermore, it is hard to control the trade-offs between low uniqueness and high complexity/sufficiency. Finally, it seems like if we have access to multiple decision cases and outcomes, it would be more helpful for auditors to just test out a range of different justification systems and compare them against each other, instead of relying solely on the justification provided by the decision-maker.

ACKNOWLEDGMENTS

This research was supported in part by the Graduate Fellowships for STEM Diversity (GFSD), as well as NSF Awards IIS-1901168 and IIS-2008139. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors. Additional thanks to Travis McGaha, Hao Tang, and Aaron Tucker for helpful comments.

REFERENCES

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2019. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. (2019). https://doi.org/10.48550/ARXIV.1910.10045 arXiv:1910.10045 [cs.AI]

[2] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. Fairwashing: the risk of rationalization. (2019). https://doi.org/10.48550/ARXIV.1901.09749 arXiv:1901.09749 [cs.LG]

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445717

[4] Solon Barocas, Andrew D. Selbst, and Manish Raghavan. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 80–89. https://doi.org/10.1145/3351095.3372830

[5] Umang Bhatt, Adrian Weller, and José M. F. Moura. 2020. Evaluating and Aggregating Feature-based Model Explanations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, 3016–3022. https://doi.org/10.24963/ijcai.2020/417

[6] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, JB Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. 2020. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. (2020). https://doi.org/10.48550/ARXIV.2004.07213 arXiv:2004.07213 [cs.CY]

[7] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (April 2021). https://doi.org/10.1145/3449287

[8] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness *(KDD '17)*. Association for Computing Machinery, New York, NY, USA, 797–806. https://doi.org/10.1145/3097983.3098095

[9] Sanjoy Dasgupta, Nave Frost, and Michal Moshkovitz. 2022. Framework for Evaluating Faithfulness of Local Explanations. (2022). https://doi.org/10.48550/ARXIV.2202.00734 arXiv:2202.00734 [cs.LG]

[10] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 275–285. https://doi.org/10.1145/3301275.3302310

[11] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. (2017). https://doi.org/10.48550/ARXIV.1702.08608 arXiv:1702.08608 [cs.AI]

[12] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Schieber, James Waldo, David Weinberger, Adrian Weller, and Alexandra Wood. 2017. Accountability of AI Under the Law: The Role of Explanation. (2017). https://doi.org/10.48550/ARXIV.1711.01134 arXiv:1711.01134 [cs.AI]

[13] Lilian Edwards and Michael Veale. 2017. Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For. *Duke Law & Technology Review* 16 (2017), 18–84. https://doi.org/10.2139/ssrn.2972855

[14] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L. Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet. Digital health* 3, 11 (Nov. 2021), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

[15] Gillian K. Hadfield. 2021. *Explanation and justification: AI decision-making, law, and the rights of citizens.* https://srinstitute.utoronto.ca/news/hadfield-justifiable-ai

[16] Ana Valeria González, Gagan Bansal, Angela Fan, Yashar Mehdad, Robin Jia, and Srinivasan Iyer. 2021. Do Explanations Help Users Detect Errors in Open-Domain QA? An Evaluation of Spoken vs. Visual Explanations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 1103–1116. https://doi.org/10.18653/v1/2021.

findings-acl.95
[17] Yoyo Tsung-Yu Hou and Malte F. Jung. 2021. Who is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021). https://doi.org/10.1145/3479864
[18] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. *Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks.* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[19] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE, San Jose, CA, USA, 3–10. https://doi.org/10.1109/VLHCC.2013.6645235
[20] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and Customizable Explanations of Black Box Models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 131–138. https://doi.org/10.1145/3306618.3314229
[21] Ruiwen Li, Zhibo Zhang, Jiani Li, Chiheb Trabelsi, Scott Sanner, Jongseong Jang, Yeonjeong Jeong, and Dongsub Shim. 2021. EDDA: Explanation-driven Data Augmentation to Improve Explanation Faithfulness. (2021). https://doi.org/10.48550/ARXIV.2105.14162 arXiv:2105.14162 [cs.AI]
[22] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376590
[23] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 2119–2128. https://doi.org/10.1145/1518701.1519023
[24] Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. *Queue* 16, 3 (June 2018), 31–57. https://doi.org/10.1145/3236386.3241340
[25] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-Distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021). https://doi.org/10.1145/3479552
[26] Hui Liu, Qingyu Yin, and William Yang Wang. 2019. Towards Explainable NLP: A Generative Explanation Framework for Text Classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 5570–5581. https://doi.org/10.18653/v1/P19-1560
[27] Shixia Liu, Xiting Wang, Mengchen Liu, and Jun Zhu. 2017. Towards Better Analysis of Machine Learning Models: A Visual Analytics Perspective. (2017). https://doi.org/10.48550/ARXIV.1702.01226 arXiv:1702.01226 [cs.LG]
[28] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6 (July 2021). https://doi.org/10.1145/3457607
[29] Tim Miller. 2017. Explanation in Artificial Intelligence: Insights from the Social Sciences. (2017). https://doi.org/10.48550/ARXIV.1706.07269 arXiv:1706.07269 [cs.AI]
[30] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems* 11, 3–4 (Aug. 2021). https://doi.org/10.1145/3387166
[31] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. (2018). https://doi.org/10.48550/ARXIV.1802.00682 arXiv:1802.00682 [cs.AI]
[32] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778
[33] Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. https://doi.org/10.48550/arXiv.1811.10154 arXiv:1811.10154 [cs, stat].
[34] Andrew Selbst and Solon Barocas. 2018. The Intuitive Appeal of Explainable Machines. *Fordham Law Review* 87, 3 (Jan. 2018), 1085. https://ir.lawnet.fordham.edu/flr/vol87/iss3/11
[35] Andrew D Selbst and Julia Powles. 2017. Meaningful information and the right to explanation. *International Data Privacy Law* 7, 4 (Dec. 2017), 233–242. https://doi.org/10.1093/idpl/ipx022 arXiv:https://academic.oup.com/idpl/article-pdf/7/4/233/22923065/ipx022.pdf
[36] Jacob Sippy, Gagan Bansal, and Daniel S. Weld. 2020. *Data Staining: A Method for Comparing Faithfulness of Explainers*. Technical Report. https://aiweb.cs.washington.edu/ai/pubs/sippy-icml20.pdf
[37] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 180–186. https://doi.org/10.1145/3375627.3375830
[38] Sandra Wachter, Brent Daniel Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 31, 2 (2017), 47. https://doi.org/10.2139/ssrn.3063289
[39] Fan Yang, Mengnan Du, and Xia Hu. 2019. Evaluating Explanation Without Ground Truth in Interpretable Machine Learning. (2019). https://doi.org/10.48550/ARXIV.1907.06831 arXiv:1907.06831 [cs.AI]
[40] Wei Zhang, Ziming Huang, Yada Zhu, Guangnan Ye, Xiaodong Cui, and Fan Zhang. 2021. On Sample Based Explanation Methods for NLP: Faithfulness, Efficiency and Semantic Evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 5399–5411. https://doi.org/10.18653/v1/2021.acl-long.419