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

Joyce Zhou[1,*], Thorsten Joachims[1]

[1]Cornell University

**Abstract**

Discussion of the legal "right to an explanation" has been increasingly relevant because of its potential usefulness in auditing automated decision systems, as well as for making objections to such decisions. However, policy proposals have been vague about what requirements such explanations would have to meet. Most past work in explainable AI has focused on explanations' potential for helping model developers or human-AI team collaboration, and less on how they may affect decision recipients. In a collaborative environment, designers are motivated to implement good-faith explanations that accurately show the weaknesses of these decision systems. In an auditing environment, this motivation may not hold. 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 be used with a black-box decision system could aim to manipulate decision recipients or auditors into failing to recognize an intentionally discriminatory decision model. We test out two scenarios: a case-by-case scenario where decision recipients are unable to share their cases and explanations with each other, and a system-wide scenario where every output could be openly shared for auditing. In the case-by-case scenario, we find that the vast majority of individual decision recipients could receive a justification that seems well grounded in data, even if the decision system is intentionally discriminatory. In the system-wide scenario, we find that while a large number of justifications might conflict with each other, there is no intuitive threshold to determine if this conflict is because of malicious justifications or because of simplicity requirements of these justifications conflicting with model behavior. We end with discussion of how explanation systems could both be useful or exploitable as audit tools.

**Keywords**

explainable AI, right to an explanation, adversarial explanations

## 1. Introduction

There's 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, as well as in generally increased calls for research and discussion in this topic [1, 2, 3]. However, these legal bills do not clarify what goals such an explanation should serve to fulfill, or what an "explanation" precisely is. What distinguishes the idea of 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 led to these discussions to start with?

Within computer science, we are only starting to understand how such explanations could affect those on the receiving end of these decisions. Furthermore, these decision recipients are much less likely to be familiar with the AI system details or have general knowledge around the decisions themselves. For instance, a criminal defendant probably doesn't know the details of how their risk assessment score was trained or how it fits into other sentencing guideline systems. How might we create an explanation of why decision(s) were made that is acceptable and/or useful to decision recipients and any potential auditors? What standards (evaluation criteria) should we hold these explanations to?

We don't answer all of these questions, but we do demonstrate that in a scenario where decision recipients don't have access to the internals of a decision model or explanation system, simply maintaining a "right to an explanation" is not enough to identify malicious decision systems. Specifically, we examine simplified data centered around COMPAS recidivism prediction and demonstrate how an opaque explanation system could be abused to defend an opaque decision system.

We imagine a situation in which the group making a decision using the outputs from this recidivism prediction model is also responsible for providing an explanation. As such, they are inclined to present explanations that defend whatever decisions they make. For clarity, we

*AIofAI '22: 2nd Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies Vienna, Austria*

*Corresponding author.

jz549@cornell.edu (J. Zhou); tj36@cornell.edu (T. Joachims)

https://cephcyn.github.io/ (J. Zhou); https://www.joachims.org (T. Joachims)

0000-0003-1205-3970 (J. Zhou); 0000-0003-3654-3683 (T. Joachims)

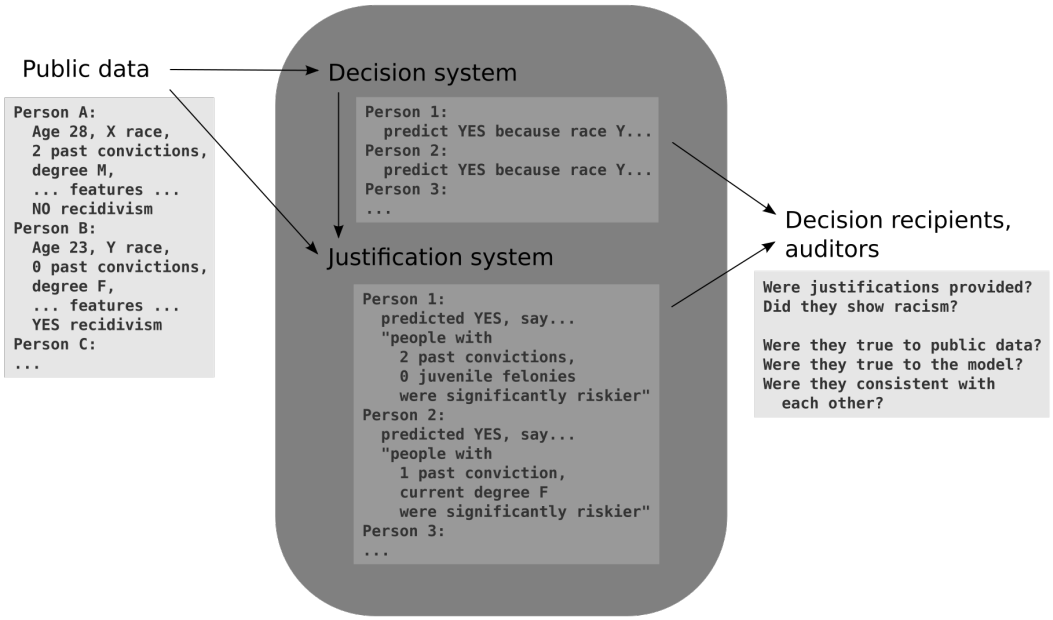


**Figure 1:** We examine scenarios where the decision-maker is responsible for fulfilling the "right to an explanation". The area with a dark grey background represents the black-box systems that are privately designed, implemented, and owned by the decision-maker. These black-box systems are designed with the intention of making decisions and defending the decision-making model as ethical. The area with a light background represents what information is publicly accessible and usable for auditing.

refer to this as a "justification" instead of an "explanation" to emphasize its purpose in justifying a decision made, in contrast to visualizing the mechanics of the decision system itself [1]. We treat both the decision model and the justification system as black boxes, where outsiders know nothing about how they actually function, since they may be trained on sensitive information or otherwise be privately owned or maintained.

We find that regardless of how accurate a decision model is, or how relevant the features visibly referenced by a justification system are to the underlying model, the majority of decision cases could be defended with a justification that appears statistically significant and supports whatever decision was made. At a simplified case-by-case level, it seems that most decisions could be defended by some kind of justification. In fact, in the majority of cases in our data both the positive and the negative decision have a seemingly valid justification. Since this case-by-case analysis ignores that decision recipients or system auditors are likely to share information, we also investigate whether shared justifications provide stronger accountability. In particular, 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 does not seem to be any intuitive faithfulness threshold that reveals whether a justification system is covering up any intentionally discriminatory decision system. In a real-world situation, it may be more effective to just audit decision systems by comparing the effectiveness of different potential justification systems using publicly accessible decision contexts and outcomes, instead of solely relying on justifications provided by a decision-maker.

## 2. Related Work

So far, most work dealing with AI interpretability or explainability has been designed by and for those working with the development and usage interfaces for AI systems. For instance, explanations may be designed for AI experts

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

and data experts who may be structuring and evaluating the model itself [4, 5], or AI novices who are end users of such systems being given assistance through AI decision-making [5]. In the context of explanations presented to experts designing and debugging a system, we might evaluate them by how well they expose biases within the system [6, 7, 5], what types of input flaws may be revealed [8, 7], or how they handle edge or adversarial cases [7, 9, 10]. For explanations presented to assist human-AI team decision-making, evaluations are frequently centered around appropriate trust [11, 12, 5], mental models [13, 5], or overall team performance [12, 5].

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 [14, 15], soundness and completeness of explanations [16], or formalizing interpretability itself and suggesting methods to evaluate it across varying model classes and tasks [17]. There has also been some focus on how the presentation of the decision model and explanations [18], or their relationships to the task [19] affect human decision-making and overall trust.

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. For example, [9] discuss how saliency maps (highlighting important areas of an image) are commonly used as explanations with medical image analysis systems, but aren't helpful 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 suggested as helping identify unethical or unacceptable AI systems [3] and providing some base to make decision objections off of [1, 2]. However, it is still unclear what requirements an explanation satisfying this right would have to satisfy [20, 21], or if providing an explanation would help these goals at all [21].

Here, we focus on the concern that manipulative explanation systems could intentionally support or defend a system. For a human decision-maker, we know they 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 [22]? [23] presents an adversarial model that could be used together with explanations based on input perturbations in order to present explanations for racist decisions that focus on innocuous features. Similarly, [24] demonstrates how unfair models could be presented with maliciously generated fair rule lists that still appear faithful to the model itself. We do something similar by exploring how simple explanations could fail to identify racist models in two different auditing scenarios.

# 3. Question: 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 might motivate 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 discussion community building up around it where decision and justification information may be 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 potential 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 a motivation 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 Criteria

We defined and identified a set of criteria for usable justifications to defend decision-making: 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 criteria, then it appears satisfactory to a decision recipient.

For a justification to be **simple**, it needs to be easily understandable. This is important, as past work [15, 14] has argued that explanations (or justifications, here) need to stay simple for people to be able to understand them. If an explanation or justification is not simple enough to understand, then it is effectively useless. 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 explanation systems for developers or decision collaborators, this is typically not a clear priority or is assumed to hold because the explanation system has direct access to model gradients [25] or decisions made on perturbed input [8]. 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.

Finally, 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.

We now describe a justification template based on these requirements. 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 [2] more or less likely to recidivate compared to the entire arrest population in this dataset. For example, one justification based on the "current charge degree" and "juvenile felony count" features 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". Because we assume a decision recipient or system auditor knows little about how the decision-making or justification system works internally, we are unable to directly compare these dataset features with the system configuration.

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 collected a dataset of COMPAS recidivism prediction cases 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.

[2]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.

## 3.2. Dataset

To build the dataset supporting model decisions and potentially usable justifications, we used the COMPAS two-year-recidivism dataset originally collected and organized by ProPublica using public records requests and public criminal records for their "Machine Bias" article [26]. We focus on recidivism prediction specifically because it greatly impacts the lives of decision recipients, but the prediction decision 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, sex, race), criminal records information (age at arrest, number of prior misdemeanors and/or felonies, relevant criminal charge at pretrial time), and COMPAS scoring information (10-point COMPAS decile score, simplification to high- vs low- risk recidivism prediction, whether the person actually recidivated within two years). Similar to ProPublica, we filter the dataset to exclude cases with charge and arrest dates not within 30 days, missing COMPAS decile scores, or were ordinary traffic offenses. 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

Each decision has a set of potentially usable (simple, verifiable, and relevant) justifications. To identify these, we abuse the multiple comparisons problem to identify every usable justification for each decision. One of these is presented to the decision recipient as a "final" justification, with its verifiability only based on one statistical significance test.

To optimize for runtime, we calculated all potentially usable justifications across the entire decision test set for any decision. Note that this does not break our assumption that each test set justification is independently calculated from the others: for each justification, we know nothing about what other decisions will need justification later.

For each case in the test set, we iterated through every possible combination of feature values for up to $N_f$ usable features, identified the subset of reference population data that matched those features, and calculated whether that reference population subset 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[3].

To test these manipulative justifications, we implemented a range of simple recidivism prediction decision systems.

## 3.4. Decision Systems

How well can our 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 is based on 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.

Finally, we describe how we measured the success rate of these justifications on each system and discuss our experiment results.

## 3.5. Justifiability Metrics

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 doesn't 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 "justifiable" at this 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 model. As it turns out, manipulative justification systems succeed at a case-by-case level in part because of how many of these universally justifiable cases there are.

## 3.6. Results: Case-by-case

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

In Figure 2, we show a sample of 10 cases each from the predicted low-risk and high-risk recidivism groups in the test population based on the original COMPAS predictions, summarized in terms of the "significant" justifications that could be applied to each of them. For the vast majority of cases, there are relevant justifications that could be used in favor of either potential prediction.

In Table 1, we count how many cases in the test set have significant justifications that could defend either type of original model decision, compared to having only justifications in favor of or against the actual predictions. In the overall test population, over 90% of cases have some justification usable in favor of the actual decision generated. Notably, *over 68% of all test set cases have applicable justifications that could be used to defend any potential decision*, regardless of the actual decision model output!

In addition, 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. If justifications are allowed to be given out case-by-case without any auditing across decisions, then the vast majority of these decisions can be defended with some kind of statistically "significant" feature-based justification.

Again, we emphasize that *all possible COMPAS decision systems based on this dataset and using this justification template have a majority of justifiable cases in the case-by-case scenario.* From Table 1, the 68% of cases that have justifications available for either decision can always be defended in a case-by-case scenario regardless of the

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

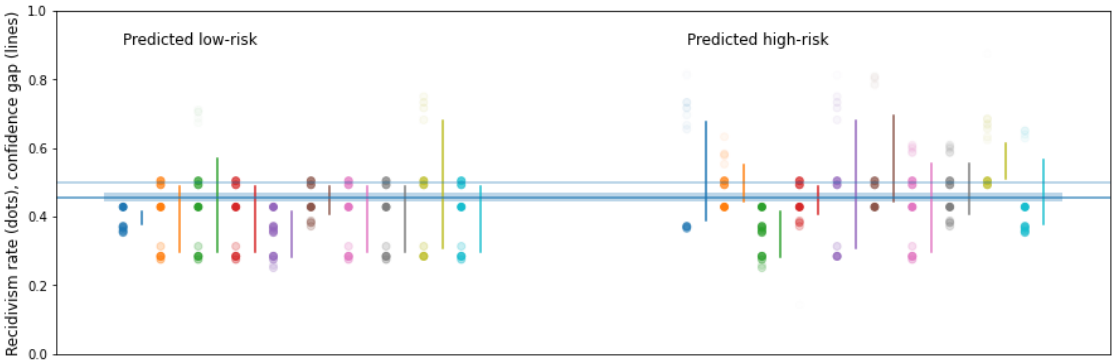


**Figure 2:** A sample of predicted low-risk and high-risk COMPAS cases and a summary of what their relevant justifications could achieve. Each case is represented by a colored group of dots and lines. Each dot associated with a case represents a unique applicable justification for that case, with its Y-value representing the mean observed recidivism rate based on the relevant reference subgroup for that justification. Each vertical line represents the gap between confidence intervals of the "least" and "most" risk decision-defending justifications for that case. The dark horizontal line with overlaid interval represents the confidence interval of the overall reference population. The lighter horizontal line is fixed at 0.5. If a vertical line extends both above and below the confidence interval, it means there are justifications that can serve towards either potential recidivism prediction. The greater the distance, the greater the confidence interval gap between population and justification subgroup.

**Table 1**
Justifiability of original model decisions, varying across justification complexity

| # Features | Neither | Both | Pro-Only | Anti-Only | % justifiable? |
|---|---|---|---|---|---|
| $N_f \leq 2$ | 0 | 846 | 284 | 105 | 91.49% |
| $N_f \leq 3$ | 0 | 850 | 281 | 104 | 91.57% |

decision model.

To demonstate this, we present a breakdown of justifiability counts across the original, racist, oracle, and random decision models in Table 2. While the percentage of justifiable cases does vary across all of these decision systems and does imply that the original decision model is more justifiable than a racist decision model, all of these are majority justifiable. The change in percentage justifiable is solely caused by the set of cases for which justifications are only available in favor of one potential decision, and what decisions are given on those cases. Note that in a real-world scenario, we may have no automatic justification generation system to compare potential upper and lower bounds for justifiability against.

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 frequently isn't the case. Decision recipients may well be able to reach out to each other and form communities. Furthermore, they may benefit from revealing if 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. Question: 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 being provided for each decision are independently provided, but all case data, decisions, and justifications are publicly visible for auditing. To do this, we run a justification-providing system that assigns each test set decision case its justification independently of every other case, and calculate global consistency, global sufficiency, and uniqueness metrics from [27]. Given the

**Table 2**
Justifiability using only $N_f \leq 2$ justifications, varying across decision models

| Model | Neither | Both | Pro-Only | Anti-Only | % justifiable? |
|---|---|---|---|---|---|
| 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% |

justifications that a justification system gives in defense of some decision system, could we assess how faithful or manipulative the justifications are?

## 4.1. Faithfulness Metrics

We used faithfulness metrics proposed by [27] 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.

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 the same justifications are being used for decisions that frequently contradict each other. There would be obvious self-conflict across multiple cases.

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 applicable by the same justification, 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. There would be implied self-conflict across multiple cases, if we know what features are shared across these cases.

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 end up being extremely unhelpful for identifying common patterns across different cases.

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

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

Global sufficiency is defined as:

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

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 $\pi$ was assigned to.
- $A(x, \pi)$ is true if justification $\pi$ 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 $\pi$ could describe.
- $\mu$ 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

We start with the assumption that a justification system is implemented to defend whatever decisions were made as best as it can. Its developers are also aware of the auditing system and faithfulness metrics, but do not know

in advance exactly what cases or decisions they will need to generate justifications for. 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 fair model that does not use protected features.

For 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 the justification system designers may be aware what metrics they are audited by, 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 conflicting with as few of the other decisions made as possible based on estimates from the reference set. 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" (a "agree-only" system).

### 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.

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, as not all decisions have an applicable defending justification. 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.

In Table 3, we compare faithfulness metrics based on what justifications would be given by the "agree-only" justification selection 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. We show this tradeoff in Table 4.

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 5.

Similar to before, these faithfulness metrics do 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 still unclear what a reasonable threshold for consistency, sufficiency, or uniqueness may be. If we only evaluate faithfulness metrics for one existing set of justifications and decisions, all but the random decision system would show high consistency, low uniqueness, and sufficiency above 0.5 (the majority of related decision cases have agreeing outcomes).

If there is no intuitive threshold we can use for justification faithfulness, how else could we use justifications to identify malicious decision systems? It is both hard and unhelpful to contrast faithfulness metrics of one justification system across different potential decision systems. Contrasting these requires that we have access to the justification system or otherwise assume how it works. Furthermore, decision systems showing higher faithfulness may have worse overall performance or otherwise over-fit to the justification system itself.

However, could we contrast the faithfulness of different potential justification systems against a known justification system and its outputs? One challenge that comes up with this approach is that we encounter issues with a uniqueness (justification complexity) vs. consistency and sufficiency trade-off: it is hard to control for justification system uniqueness when we allow justifications with

**Table 3**
System-wide faithfulness using only $N_f \leq 2$ "agree-only" justifications, varying across decision models

| Model | % Justified (↑) | Consistency (↑) | Sufficiency (↑) | Uniqueness (↓) |
|---|---|---|---|---|
| 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 |

**Table 4**
System-wide faithfulness using only $N_f \leq 3$ "agree-only" justifications, varying across decision models

| Model | % Justified (↑) | Consistency (↑) | Sufficiency (↑) | Uniqueness (↓) |
|---|---|---|---|---|
| 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 |

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.

We contrast faithfulness metrics from this extended justification with those of the original in Table 6. We can see a huge gap between race-using and race-excluding justification faithfulness for the "racist" decision model, with sufficiency especially increasingly 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 relatively simple justifications and evaluate them.

## 5. Discussion

In the vast majority of test cases, it is possible to provide a poor-faith justification at a case-by-case level that still appears simple and verifiable for a recividism prediction decision by taking advantage of the multiple comparisons problem. For a smaller but critically important majority, it is possible to do this in favor of either potential prediction: whether a manipulative justifier is defending a high-risk or low-risk recidivism predictions, there is a cherry-picked statistical comparison available for them to use. On a dataset with more fields, we speculate that the percentage of justifiable cases would only increase.

Overall, the right to an explanation could easily be

**Table 5**
System-wide faithfulness using only "must-justify" justifications, varying across decision models

| # Features | Model | % Justified (↑) | Consistency (↑) | Sufficiency (↑) | Uniqueness (↓) |
|---|---|---|---|---|---|
| $N_f \leq 2$ | 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 |

**Table 6**
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 |

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 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 [15, 14], soundness and completeness [16], 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, this added complexity 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 most obvious 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 seems to indicate that the "right to an explanation" is not directly helpful for verifying the validity of decision systems alone. If we are checking for overall system validity, it seems more effective to enable full audits from outside observers with accessible decision outputs.

While there may still be ways to make use of explanation systems in automated decision-making, such as highlighting decision feedback or adjustment mechanics [2], using them for system auditing while relying on the decision provider to give a justification does not seem like a trustworthy or reliable way to do that. Instead, open communication and third-party examination across multiple cases seem more effective for system auditing.

In the future, it might also be interesting to explore different explanation system quality metrics. 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 [9], but could we also use it as an explanation system evaluation metric?

## 6. 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 justifica-

tion 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 tradeoff between uniqueness/complexity and overall faithfulness metrics. 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 for helpful comments.

## References

[1] A. D. Selbst, J. Powles, Meaningful information and the right to explanation, International Data Privacy Law 7 (2017) 233–242. URL: https://doi.org/10.1093/idpl/ipx022. doi:10.1093/idpl/ipx022, _eprint: https://academic.oup.com/idpl/article-pdf/7/4/233/22923065/ipx022.pdf.

[2] S. Wachter, B. D. Mittelstadt, C. Russell, Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR, Cybersecurity (2017). URL: https://api.semanticscholar.org/CorpusID:3995299. doi:https://dx.doi.org/10.2139/ssrn.3063289.

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

[4] S. Liu, X. Wang, M. Liu, J. Zhu, Towards better analysis of machine learning models: A visual analytics perspective, Vis. Informatics 1 (2017) 48–56. URL: https://api.semanticscholar.org/CorpusID:7545060.

[5] S. Mohseni, N. Zarei, E. D. Ragan, A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems, ACM Trans. Interact. Intell. Syst. 11 (2021). URL: https://api.semanticscholar.org/CorpusID:208910731. doi:10.1145/3387166, place: New York, NY, USA Publisher: Association for Computing Machinery.

[6] J. Dodge, Q. V. Liao, Y. Zhang, R. K. E. Bellamy, C. Dugan, Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment, in: Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 275–285. URL: https://api.semanticscholar.org/CorpusID:59158762. doi:10.1145/3301275.3302310, event-place: Marina del Ray, California.

[7] A. B. Arrieta, N. D'iaz-Rodr'iguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garc'ia, S. Gil-L'opez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI, ArXiv abs/1910.10045 (2020). URL: https://api.semanticscholar.org/CorpusID:204824113.

[8] M. T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016). URL: https://api.semanticscholar.org/CorpusID:13029170.

[9] M. Ghassemi, L. Oakden-Rayner, A. Beam, The false hope of current approaches to explainable artificial intelligence in health care., The Lancet. Digital health 3 11 (2021) e745–e750. URL: https://api.semanticscholar.org/CorpusID:239963176. doi:10.1016/S2589-7500(21)00208-9.

[10] H. Liu, V. Lai, C. Tan, Understanding the Effect of Out-of-Distribution Examples and Interactive Explanations on Human-AI Decision Making, Proc. ACM Hum.-Comput. Interact. 5 (2021). URL: https://api.semanticscholar.org/CorpusID:231603356. doi:10.1145/3479552, place: New York, NY, USA Publisher: Association for Computing Machinery.

[11] A. V. Gonzalez, G. Bansal, A. Fan, Y. Mehdad, R. Jia, S. Iyer, Do Explanations Help Users Detect Errors in Open-Domain QA? An Evaluation of Spoken vs. Visual Explanations, in: FINDINGS, 2021. URL: https://api.semanticscholar.org/CorpusID:236478213.

[12] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Ka-

mar, M. T. Ribeiro, D. Weld, 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, Association for Computing Machinery, New York, NY, USA, 2021. URL: https://api.semanticscholar.org/CorpusID:220128138.

[13] B. Y. Lim, A. K. Dey, D. Avrahami, 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, CHI '09, Association for Computing Machinery, New York, NY, USA, 2009, pp. 2119–2128. URL: https://api.semanticscholar.org/CorpusID:4507550. doi:10.1145/1518701.1519023, event-place: Boston, MA, USA.

[14] M. Narayanan, E. Chen, J. He, B. Kim, S. J. Gershman, F. Doshi-Velez, How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation, ArXiv abs/1902.00006 (2018). URL: https://api.semanticscholar.org/CorpusID:3652280.

[15] T. Miller, Explanation in Artificial Intelligence: Insights from the Social Sciences, Artif. Intell. 267 (2019) 1–38. URL: https://api.semanticscholar.org/CorpusID:36024272.

[16] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, W.-K. Wong, 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, 2013, pp. 3–10. URL: https://api.semanticscholar.org/CorpusID:6960803. doi:10.1109/VLHCC.2013.6645235.

[17] F. Doshi-Velez, B. Kim, Towards A Rigorous Science of Interpretable Machine Learning, arXiv: Machine Learning (2017). URL: https://api.semanticscholar.org/CorpusID:11319376.

[18] Z. Buçinca, M. B. Malaya, K. Z. Gajos, To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making, Proc. ACM Hum.-Comput. Interact. 5 (2021). URL: https://api.semanticscholar.org/CorpusID:231979279. doi:10.1145/3449287, place: New York, NY, USA Publisher: Association for Computing Machinery.

[19] Y. T.-Y. Hou, M. F. Jung, Who is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making, Proc. ACM Hum.-Comput. Interact. 5 (2021). URL: https://api.semanticscholar.org/CorpusID:239020696. doi:10.1145/3479864, place: New York, NY, USA Publisher: Association for Computing Machinery.

[20] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. J. Gershman, D. O'Brien, S. Schieber, J. Waldo, D. Weinberger, A. Wood, Accountability of AI Under the Law: The Role of Explanation, ArXiv abs/1711.01134 (2017). URL: https://api.semanticscholar.org/CorpusID:2092882.

[21] L. Edwards, M. Veale, Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For, Duke law and technology review 16 (2017) 18–84. URL: https://api.semanticscholar.org/CorpusID:59148202, https://strathprints.strath.ac.uk/61618/8/Edwards_Veale_DLTR_2017_Slave_to_the_algorithm_why_a_right_to_an_explanation_is_probably.pdf.

[22] Z. C. Lipton, The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery., Queue 16 (2018) 31–57. URL: https://api.semanticscholar.org/CorpusID:5981909. doi:10.1145/3236386.3241340, place: New York, NY, USA Publisher: Association for Computing Machinery.

[23] D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju, Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, 2020, pp. 180–186. URL: https://doi.org/10.1145/3375627.3375830.

[24] U. Aïvodji, H. Arai, O. Fortineau, S. Gambs, S. Hara, A. Tapp, Fairwashing: the risk of rationalization, 2019. URL: https://arxiv.org/abs/1901.09749. doi:10.48550/ARXIV.1901.09749.

[25] H. Liu, Q. Yin, W. Y. Wang, 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, 2019, pp. 5570–5581. URL: https://api.semanticscholar.org/CorpusID:53153643. doi:10.18653/v1/P19-1560.

[26] Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner, Machine Bias (2016). URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[27] S. Dasgupta, N. Frost, M. Moshkovitz, Framework for Evaluating Faithfulness of Local Explanations, ArXiv abs/2202.00734 (2022). URL: https://api.semanticscholar.org/CorpusID:246473190.