Proceedings of the Tenth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2022)

"If it didn't happen, why would I change my decision?": How Judges Respond to Counterfactual Explanations for the Public Safety Assessment

Yaniv Yacoby,[1] Ben Green,[2] Christopher L. Griffin, Jr.,[3] Finale Doshi-Velez[1]

[1] Harvard University
[2] University of Michigan
[3] James E. Rogers College of Law, University of Arizona
yanivyacoby@g.harvard.edu, bzgreen@umich.edu, chrisgriffin@arizona.edu, finale@seas.harvard.edu

Abstract

Many researchers and policymakers have expressed excitement about algorithmic explanations enabling more fair and responsible decision-making. However, recent experimental studies have found that explanations do not always improve human use of algorithmic advice. In this study, we shed light on how people interpret and respond to counterfactual explanations (CFEs)—explanations that show how a model's output would change with marginal changes to its input(s)—in the context of pretrial risk assessment instruments (PRAIs). We ran think-aloud trials with eight sitting U.S. state court judges, providing them with recommendations from a PRAI that includes CFEs. We found that the CFEs did not alter the judges' decisions. At first, judges misinterpreted the counterfactuals as real—rather than hypothetical—changes to defendants. Once judges understood what the counterfactuals meant, they ignored them, stating their role is only to make decisions regarding the actual defendant in question. The judges also expressed a mix of reasons for ignoring or following the advice of the PRAI without CFEs. These results add to the literature detailing the unexpected ways in which people respond to algorithms and explanations. They also highlight new challenges associated with improving human-algorithm collaborations through explanations.

1 Introduction

Many high-stakes decisions are now made by people with the aid of algorithmic advice, including in the labor market and the criminal justice system. In order to promote more informed and responsible human uses of algorithms, many researchers, policymakers, and engineers have expressed interest in explainable AI (XAI). The goal of XAI systems is for algorithms to provide reasons behind the recommendations they generate. In theory, such systems could help human decision-makers identify errors, correct for biases, and synthesize the system's reasoning with their own (e.g., Ribeiro, Singh, and Guestrin (2016); Lundberg and Lee (2017); Adler et al. (2018); Bach et al. (2015); Guidotti et al. (2018)). This has motivated regulatory bodies to recommend explanations for high-stakes AI systems (e.g., Engstrom et al. (2020); High-Level Expert Group on AI (2019); UK Information Commissioner's Office (2020)).

Alongside this excitement about XAI, however, recent work has demonstrated that these benefits are not always realized in practice. User studies in multiple contexts have found that explanations did not help people evaluate the quality of algorithmic advice or incorporate that advice into decisions (e.g., Lai and Tan (2019); Bansal et al. (2021); Jacobs et al. (2021); Green and Chen (2019b)). Thus, it is not yet clear whether and under what conditions explanations improve human-algorithm decision-making. Moreover, most prior user studies have been limited to laypeople. It is particularly important to understand how practitioners interact with algorithmic explanations.

In this study, we investigated how sitting U.S. state court judges interact with a type of explanation known as "counterfactual explanations" (CFEs). CFEs provide a human decision-maker with information about how a model's output changes based on variations in the input(s) (e.g., Martens and Provost (2014); Wachter, Mittelstadt, and Russell (2018); Goyal et al. (2019); Verma, Dickerson, and Hines (2020); Stepin et al. (2021)). In theory, this information could provide human decision-makers with a better understanding of how sensitive—or robust—the model is to marginal changes to its input(s). This insight could then help decision-makers determine whether and how to use the model's predictions.

We focused on CFEs for algorithmic pretrial risk assessment instruments (PRAIs). PRAIs take information about a defendant (e.g., age, prior failures to appear) as inputs and calculate one or more risk scores (e.g., the defendant's likelihood of being arrested if released pretrial) to guide pretrial detention decisions (e.g., whether to release or detain the defendant). Although PRAIs are often touted as mitigating human biases, the underlying models themselves have been shown to perpetuate systematic biases and reproduce structural inequities (e.g., Angwin et al. (2016); Green (2020); Koepke and Robinson (2018); Barabas et al. (2018)). Given these concerns, PRAIs present a salient application in which policymakers and scholars focus on XAI (Green 2022).

We investigated whether CFEs alter how judges understand and respond to a PRAI. For example, consider a pretrial defendant who was labeled "high risk" by a PRAI that provides CFEs. A CFE might reveal that this classification was a function of a prior felony conviction (a disproportionate outcome for non-white defendants): had the de-

fendant not been convicted, they would have been deemed "low risk." Would such information—instead of just the risk score—encourage a judge to inquire about the charge behind the felony conviction? If so, might they lower their risk estimate if the charge were for a non-violent crime? Alternatively, suppose the CFE revealed that marginal changes to a defendant's criminal history significantly would change the defendant's risk classification; sometimes up, sometimes down. Would this information encourage a judge to deem the risk assessment less reliable overall? Furthermore, would any of these reactions be swayed by (perhaps implicit) bias about the defendant's race, only exacerbating existing systemic bias?

To explore how judges respond to CFEs, we conducted qualitative think-aloud studies with eight sitting judges in two U.S. states that use a PRAI called the Public Safety Assessment (PSA). We presented judges with hypothetical pretrial cases, along with the PSA report and CFEs, and prompted them to share their reasoning in real time. This provided us with detailed knowledge about how judges reason about PRAIs and CFEs.

The judges responded to the CFEs in both unexpected—and unexpectedly consistent—ways. Initially, the judges were unsure about how to interpret the CFEs, assuming that they presented factual changes to the defendants' profiles. Once they understood that the explanations represented hypothetical changes to the defendants' profiles, they ignored them, claiming the need to focus on the defendant at hand. They did not consider what the explanations signaled about the model's sensitivity. This behavior persisted even after we provided explicit coaching about what information the explanations contained and how it could be used.

We hypothesize that judges had difficulty understanding the CFEs not because they lack training in statistics or AI, but because of their legal training. Judges are trained to think counterfactually—but only about the facts of the case before them, not facts about the defendant. When considering whether a defendant caused a victim's injury, for example, a judge might naturally ask whether the injury would have occurred without the defendant's actions. A judge will not question whether the victim might have been uninjured had the defendant been a different person who had different interactions with the criminal justice system before encountering the victim. Thus, the counterfactual reasoning judges employ focuses on particular actions of the *defendant*, whereas CFEs provide insight about the reasoning of the *model*.

Our findings contribute to a growing body of literature suggesting that people react to explanations in often unexpected ways. While our study sample was relatively small, the use of practitioners as participants and the consistency of the judges' reactions provide new insights into the challenges associated with achieving any benefits from CFEs. The full version of this paper, including appendix, can be found at https://arxiv.org/abs/2205.05424.

2 Related Work

Pretrial Risk Assessment. PRAIs have gained traction as a tool for criminal justice reform efforts, but they suffer from a variety of flaws. PRAIs are typically hailed for replacing biased human judgments with "objective" algorithmic judgments (e.g., Arnold Ventures (2019a); New Jersey Courts (2017)), enabling more consistent and fair decision-making. However, PRAIs depend on data that reflect human biases and systemic inequities, thus perpetuating these biases and inequities (e.g., Julia Angwin et al. (2016); Barabas et al. (2018); Green (2020); Koepke and Robinson (2018)). Furthermore, rather than eliminate human discretion, PRAIs shift discretion to different actors and decision points (Green 2020). Broadly, because PRAIs legitimize policies associated with mass incarceration, critics have challenged their ability to promote decarceral criminal justice reform (Green 2020; Koepke and Robinson 2018; Barabas et al. 2018).

The interaction between PRAIs and human decision-makers poses another significant challenge for criminal justice reform. Experimental research demonstrates that laypeople respond to PRAIs in a racially biased manner, being more susceptible to increasing risk estimates following a risk assessment's outputs when evaluating Black defendants relative to white defendants (Green and Chen 2019a,b). Similar behaviors seem to manifest in practice. Judges across U.S. jurisdictions frequently override the PRAI's outputs that recommend releasing defendants, leading to much higher pretrial detention rates than expected when risk assessments are used (Human Rights Watch 2017; Sheriff's Justice Institute 2016; Steinhart 2006; Stevenson 2018; Stevenson and Doleac 2021). Furthermore, counter to expectations, the use of PRAIs has increased racial disparities because judges respond to recommendations in more punitive ways when evaluating Black as opposed to white defendants Albright (2019); Cowgill (2018); Stevenson and Doleac (2021).

An important limitation of prior work on human interactions with PRAIs is a lack of knowledge about the thought processes that criminal justice practitioners follow when incorporating PRAI recommendations into their decisions. Earlier studies have primarily relied on lab experiments with laypeople (e.g., Green and Chen (2019a,b); Grgić-Hlača, Engel, and Gummadi (2019)) or empirical analyses about judicial decisions using observational data (e.g., Albright (2019); Stevenson (2018)). Thus, with a few exceptions (Brayne and Christin 2020; Hannah-Moffat, Maurutto, and Turnbull 2009), we have little understanding of the thought processes that criminal justice practitioners follow when incorporating PRAI advice into their decisions. Our study contributes to research on the implementation of PRAIs by leveraging detailed, real-time descriptions of how judges reason about these models.

Explanations and Their Challenges. Previous work studying how people react to algorithmic explanations demonstrates that different ways of presenting explanations can alter how people perceive the fairness of those decisions (e.g., Binns et al. (2018); Dodge et al. (2019)). Furthermore, recent work shows that explanations do not always improve people's ability to effectively use algorithmic predictions (e.g., Bansal et al. (2021); Green and Chen (2019b)). Explanations can increase human trust in the algorithm's recom-

mendations, even when these recommendations are incorrect (e.g., Bansal et al. (2021); Jacobs et al. (2021)) or when the explanations do not accurately represent the algorithm's inner workings (e.g., Lai and Tan (2019)). Our work adds to this literature by examining how practitioners respond to explanations in a specific, high-stakes (albeit artificial) setting: judges making pretrial release decisions.

Counterfactual Explanations (CFEs). A CFE is a statement of the form, "if input x would have been altered by a small change (now called x'), a decision making system would output y' instead of y" (Wachter, Mittelstadt, and Russell 2018). In this statement, x, y are the "factual" input and output and x', y' are the "counterfactual" input and output. Formally, CFEs are the solutions to an optimization problem of the form,

$$\text{argmin}_{x'} d(x, x') \quad \text{subject to} \quad f(x') = y', \qquad (1)$$

where $d(\cdot, \cdot)$ is a distance metric, $f(x')$ is the output of the decision making system, and y' is the desired outcome (Verma, Dickerson, and Hines 2020). Looking at the set of inputs for which the decision making system outputs y' (i.e. all x' for which $f(x') = y'$), the above optimization problem selects the x' that is closest to x (according to distance metric $d(\cdot, \cdot)$). By framing the problem in this way, we find the "minimal" change from x to x' that will change the outcome from y to y'. Counterfactuals generated in this way follow guidelines from psychology (e.g., Miller, Howe, and Sonenberg (2017); Keane et al. (2021)) to be maximally intuitive by ensuring they are both "sparse" (requiring few features to change) and "proximate" (close to the original factual).

A growing body of literature (e.g., Miller, Howe, and Sonenberg (2017); Wachter, Mittelstadt, and Russell (2018); Byrne (2019); Mueller et al. (2019)) calls for the use of CFEs due to the importance of counterfactual reasoning in human thought, as historically demonstrated in fields such as psychology (e.g., Kahneman and Tversky (1981); Roese (1997); McCloy and Byrne (2000); Hilton, McClure, and Slugoski (2005); Byrne (2007, 2016)) and philosophy (e.g., Woodward (2005); Lewis (2013)), and because CFEs potentially comply with emerging regulations governing the use of AI (Wachter, Mittelstadt, and Russell 2018).

Despite this growing interest in CFEs, we do not yet know whether, when, and how CFEs enable people to make more accurate and fair decisions given algorithmic advice. Notably, two recent surveys of CFEs call out a significant lack of user studies testing how people respond to CFEs (Verma, Dickerson, and Hines 2020; Keane et al. 2021). Our study adds to the literature by providing evidence about how practitioners interpret and respond to CFEs when reasoning about high-stakes decisions in a laboratory setting.

3 The Public Safety Assessment (PSA)

In this section, we provide background on current PRAIs and how they are integrated into pretrial release decision-making.

The Public Safety Assessment (PSA) Algorithm. The PSA is a PRAI developed by Arnold Ventures to "provide judges with objective and consistent data to make informed decisions" about pretrial defendants (Arnold Ventures 2019b). The PSA takes static information about the defendant's age, the pending charge, and seven measures of prior criminal justice experiences as inputs. It assigns these risk factors an initial set of integer weights, which are then converted into two risk scores—the New Criminal Activity (NCA) and Failure to Appear (FTA) scores. These risk scores range from 1 to 6, with larger values indicating higher risk. The NCA score reflects the risk that the defendant will be arrested for a new charge if they are released, and the FTA score reflects the risk for the defendant's failure to appear to a subsequent hearing if they are released. Lastly, the new violent criminal activity (NVCA) flag provides a binary indicator that the risk of arrest for a new violent charge is high. These scores are presented to judges before or during pretrial hearings, at which judges must decide whether to release (with or without conditions, including money bail) or detain a defendant before trial.

FTA \ NCA	1	2	3	4	5	6
1	ROR	ROR				
2	ROR	ROR	PML1	PML2	PML3	
3		PML1	PML1	PML2	PML3	No Release
4		PML1	PML1	PML2	PML3	No Release
5		PML2	PML2	PML3	PML3+EM/HD	No Release
6				No Release	No Release	No Release

Figure 1: Reproduction of DMF Matrix from the New Jersey Court System (2018). This matrix maps the risk scores generated by the PSA—the FTA and NCA scores—to a color-coded release recommendation ranging from least restrictive (green) to most restrictive (red). We note that the matrix is also accompanied by additional instructions depending on the charges filed. For example, for a murder or felony-murder charge, the instructions will recommend "no release" regardless of the PSA's outputs, (https://www.njcourts.gov/courts/assets/criminal/decmakframwork.pdf). Key: ROR = release on their own recognizance; PML = Pretrial Monitoring Level; EM = Electronic Monitoring; HD = Home Detention.

The Decision-Making Framework (DMF). The NCA and FTA risk scores generated by the PSA (and the NVCA flag) are inputs into the Decision-Making Framework (DMF). In short, the DMF converts the three indicators into a recommendation for release conditions, if any. The central component of the DMF is the DMF Matrix, which combines the NCA and FTA scores to yield recommended release conditions (see Figure 1). Recommendations can diverge from a jurisdiction's DMF Matrix based on the type of charge and the NVCA flag. Local stakeholders (e.g., the court, public defenders, prosecutors) collectively choose the recommendations corresponding to combinations of NCA and FTA scores. These choices usually reflect policy goals and values in the jurisdiction, statutory requirements, and available resources (e.g., pretrial services). Thus, while the PSA scores in all jurisdictions rely on the same data and are

the product of the same scoring algorithm, the DMF varies across jurisdictions. Finally, decisions about releasing or detaining defendants (and under what conditions) always remain subject to judicial discretion. In this work, when referring to the "model," we mean the combined PSA and DMF system (PSA-DMF).

4 Methodology

We designed a survey to assess how judges respond to PSA-DMF recommendations and to CFEs of those recommendations. In the main body of the survey, we asked participating judges to make release decisions for synthetic defendants. Judges were first asked to make these decisions based on factual information about defendants (e.g., age, criminal history) and the PSA-DMF reports. Judges were then presented with CFEs (as additional information) and had the opportunity to update or stand by their initial decision.

We will first describe how we generated defendant data, DMF recommendations, and CFEs. We will then describe the participants we recruited and the experimental protocols.

4.1 Experiment Setup

Defendant Data. We created a set of realistic cases based on a sample of 500 de-identified cases from a county in Iowa that had used the PSA-DMF, consisting of actual defendants' statutory charges, risk factor values, and PSA-DMF recommendations. To ensure judges were not presented with a homogeneous set of cases, we selected a subset of cases with a diverse set of common charges, and for which the CFEs included a variety of hypothetical changes leading to more or less restrictive recommendations (or both). Lastly, we synthesized the demographic information, including the date of birth, date of arrest, race, and name (ensuring the name was likely perceived to correspond to the defendant's race, using the top first names identified by Gaddis (2017) and top last names identified by Word et al. (2008)). See Appendix A.1 for more details, and see Figure 2 for an example defendant.

Generating DMF Recommendations. We used a DMF matrix similar to the one in Figure 1. Since we generated counterfactuals of the combined PSA-DMF relative to the PSA's *inputs* (see below), we chose not to include a full visualization of the DMF matrix, which illustrates how the DMF recommendations change relative to the PSA's *outputs*. Instead of using a matrix of colors and acronyms for the release options (e.g., ROR, PML1, etc.), we assigned the different release options letters A through F (from least to most restrictive) and presented them as a list—see Figure 3.

Generating Counterfactuals. Following the guidelines from Section 2, we designed the counterfactuals to be intuitive by ensuring they were both sparse and proximate. For each defendant in the cohort, we iterated over all possible risk factor combinations (with only nine factors, one can easily iterate through all of them), keeping only risk factor combinations that were (a) consistent with one another, (b) a single "edit" different from the original defendant, and (c) for which the model's final recommendation differs from the original recommendation. Example counterfactuals are visualized in Figure 4; we chose this representation for its concision relative to other types of visualizations (e.g., heatmaps).

4.2 Think-Aloud Protocols

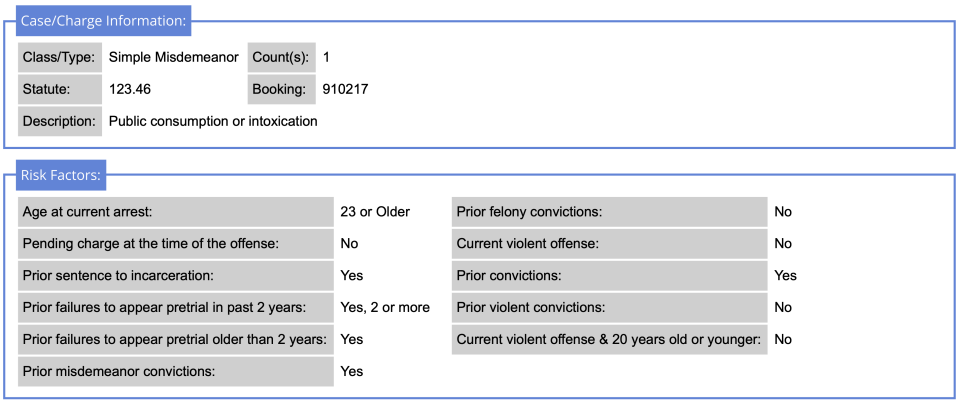
Participants. We recruited eight judges to participate in our study—four judges from a Mountain region state for Round 1 and four judges from a Southwest region state for Round 2—via direct solicitation and snowball sampling. All participants were judges in districts that use the Arnold Ventures PSA-DMF. Thus, all participants already had regular experience with the model, to which our study added the CFEs. Participants had between 3 and 26.5 years of experience on the bench (mean of 16.4 years).

Study Overview. The think-aloud protocol for both rounds of the survey are included in Appendices A and B; screenshots of the survey are included in Appendix D, along with example cases in Appendix E. After obtaining consent, participants received a short tutorial on our adapted PSA-DMF report (Figures D.1 and D.2), including what CFEs are and how to read them. Next, to ensure participants understood the information presented in the CFEs, we asked basic comprehension questions about the CFE (e.g., "If the given defendant had one less prior conviction, what would the model recommend?")—see Appendix C for details.

After the consent process, instructions, and comprehension questions, we moved to the main body of the survey (Figures D.6 to D.9). We presented each synthetic defendant's risk factors, demographic information, and the model's release recommendation (without any explanation). We then asked participants to narrate their thought processes while reading and interpreting the PSA-DMF report before arriving at their release decision. After formulating that release decision, we showed them the CFE and asked them whether they wanted to revise their original decision based on the new information (and if so, why). For those who were hesitant to think aloud or who stopped sharing in the middle of the survey, we prompted them to continue by asking questions such as, "What led you to choose that option?"

For visual distinction, we placed all model outputs in red boxes, i.e., the FTA and NCA scores and DMF recommendations, as well as the CFEs (Figures 3 and 4). We placed all remaining information, i.e., the inputs to the model (risk factors), charge and demographic information, in blue boxes (Figure 2). We note the colors here because some judges referred to the boxes by color.

We conducted all interviews via Zoom and limited them to about one hour to respect the judges' time. Judges completed an average of 4.4 of the cases in that time, and, with their permission, we recorded every interview. We had each interview professionally transcribed. Two members of the research team then reviewed these transcripts to characterize the views expressed by judges during the interviews. We present each finding with quoted language from multiple participants in order to demonstrate the common threads across responses. This work was approved by the [Anonymous] Institutional Review Board.



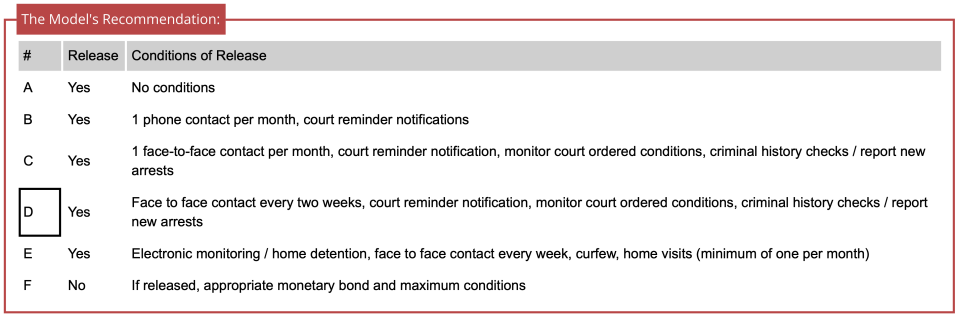
Case/Charge Information:

Class/Type:	Simple Misdemeanor	Count(s):	1
Statute:	123.46	Booking:	910217
Description:	Public consumption or intoxication		

Risk Factors:

Age at current arrest:	23 or Older	Prior felony convictions:	No
Pending charge at the time of the offense:	No	Current violent offense:	No
Prior sentence to incarceration:	Yes	Prior convictions:	Yes
Prior failures to appear pretrial in past 2 years:	Yes, 2 or more	Prior violent convictions:	No
Prior failures to appear pretrial older than 2 years:	Yes	Current violent offense & 20 years old or younger:	No
Prior misdemeanor convictions:	Yes		

Figure 2: Example case/charge information and risk factors from our survey.



The Model's Recommendation:

#	Release	Conditions of Release
A	Yes	No conditions
B	Yes	1 phone contact per month, court reminder notifications
C	Yes	1 face-to-face contact per month, court reminder notification, monitor court ordered conditions, criminal history checks / report new arrests
D	Yes	Face to face contact every two weeks, court reminder notification, monitor court ordered conditions, criminal history checks / report new arrests
E	Yes	Electronic monitoring / home detention, face to face contact every week, curfew, home visits (minimum of one per month)
F	No	If released, appropriate monetary bond and maximum conditions

Figure 3: Example DMF recommendations from our survey. In this example, the model (PSA-DMF) recommends option 'D', which is for the defendant to be released, but with face-to-face contact every two weeks, etc.

Differences between Rounds 1 and 2. When conducting interviews with the first four judges (Round 1), we found surprising yet consistent results. As we describe in more detail in Section 5.2, even with the tutorial and comprehension questions, the judges interpreted CFEs as new *factual* information: they thought that the actual defendant's profile had been updated. Once they understood that the CFEs contained hypothetical scenarios, the judges still largely ignored them. They believed it inappropriate to consider hypothetical information about a defendant when making decisions about release—even though the hypotheticals were about the PSA-DMF outputs. In other words, the judges mistakenly confused the CFEs as indicative of a new defendant before them, as opposed to new information about the functioning of the PRAI using the same defendant.

These findings prompted us to adjust our protocol. In Round 1, we deliberately excluded specific training on how the judges might use the counterfactuals to adjust their decisions, because we did not want to influence their behavior. However, our initial results prompted us to ask whether judges would respond differently if we provided more training on what the counterfactuals meant and how judges might use them. Although doing so might have primed the judges to use the explanation in certain ways, it would also help us explore whether the results in Round 1 were due to a lack of understanding of our instructions or actual behavior.

For Round 2 (detailed in Appendix B), we therefore expanded our training to provide more information about the CFEs. We created a presentation to describe what a CFE was and concrete ways in which it could be used (e.g., if the defendant's age is 22 or 23 years, the judge can use the CFE to "adjust" the final recommendation, depending on whether they believe that age is an important factor). In the presentation, we additionally asked judges to explain to us what the CFE was to ensure they understood it. We hoped that having a live, interactive presentation at the start of the survey would ensure that participants paid attention (rather than potentially skimmed the tutorial) and felt comfortable asking clarifying questions (as opposed to feeling self-conscious about their reading and comprehension speed). Indeed, the judges felt more comfortable asking questions; when asked to explain the counterfactuals to us, nearly all of them did so well.

5 Results

Across all eight think-aloud participants (four in each of the two rounds), we observed notable consistency in how the judges used the PSA-DMF and CFEs. We therefore present the results from both rounds jointly. As a reminder, when the judges refer to the "red boxes," they are referring to information based on the output of the PSA-DMF (or model), and when they refer to the "blue boxes," they refer to all other information, including the inputs to the PSA-DMF.

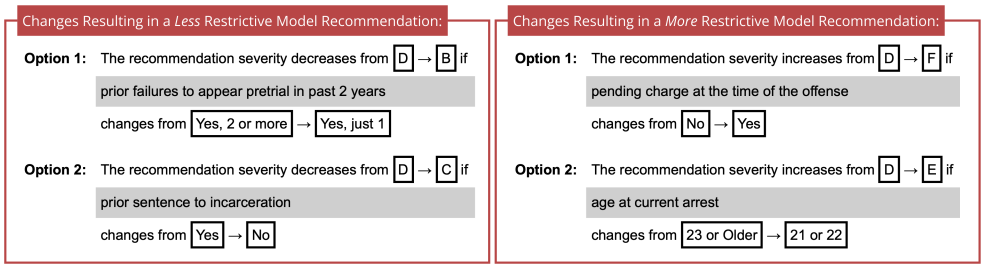


Figure 4: Example CFE from our survey. The two red boxes present single changes to the defendant's profile that would decrease and increase, respectively, how restrictive the release recommendation is. Inside each of the red boxes are several options, each of which describes the single change to the defendant's profile that would change the model's recommendation.

5.1 How Judges Use the PSA-DMF

Because we first asked judges to evaluate each defendant with the PSA-DMF report but without the CFEs, our think-aloud interviews provided insights into how judges reason when making PRAI-assisted pretrial release decisions. Judges' responses to the model's information revealed mixed attitudes regarding the utility and quality of the PSA-DMF.

Judges focused on the specific attributes of defendants, rather than the risk scores. When thinking aloud about release decisions, judges focused first on the charge class, i.e., whether the charge was a felony or misdemeanor, and the violence flag, noting that the charge class helped them determine whether the defendant is a risk to society. Next, they scanned the risk factors only to check if they were unusual or inconsistent with other information; otherwise, they ignored them and continued to the DMF's release recommendation.

> "I kind of pretty much disregarded the risk scores. [...] I gave the risk assessment a glance, but I was more interested in the detail of the risk factors. [...] Many judges want to see the details of the thing as opposed to focusing on in what we're calling here the risk assessment itself. It's of interest, but the blue box is more interesting to me." – P1

> "I mostly relied on the blue boxes, less on the red boxes." – P2

> "The risk factor box is the one that's frankly the most important to me. And, what I'm looking at and what is very important to me when I look at this is typically pending charges at the time of the offense, failures to appear, the two failure to appear boxes, obviously the violent offense boxes. These are the ones that are most important to me. Secondarily to that is probably the sort of where you land on the matrix [the FTA and NCA scores]. " – P4

> "I'm looking at the kind of charge it is, I'm looking at the PSA. I'm looking at whether there's a victim, whether it's a violent offense, whether somebody died. If it's a violent offense, the more likely it is that they're going to be held. If it's a sexual offense, and I see a lot of PSAs where they recommend release, personally they're probably going to be held [on a high bond] despite the recommendation." – P5

> "To me, a big part of my decision making is based on the facts that are alleged in the case." – P7

Some judges expressed concerns about the accuracy of the risk scores associated with the PSA model. In particular, judges expressed concerns about specific cases, for example those involving domestic violence charges.

> "I never look at these [PSA recommendations], honestly, when I'm making decisions in court. [...] I know that the [colleagues] I've talked to, we don't spend a lot of time looking at the PSA. [...] Sometimes the recommendations strike us as very off, given what other information we have. [...] It's not something I don't think any of us spend the majority of our time on, when we're making these decisions. [...] I feel like the PSA is totally, totally not valid for domestic violence cases." – P7

> "[Given a defendant with no previous criminal history, but with age $\leq$ 20:] This is actually one of the things I disagree with on PSA because it tells me new criminal activity score. And he's got him for a[n NCA of] two, but he's only got the one charge. So how would you have a one [charge], if you're saying it's new criminal activity based on the actual charge you're seeing the guy on." – P8

Here, the judge is concerned that a defendant with no prior criminal history is given an NCA of 2 (likely because younger defendants are empirically at higher risk of being re-arrested after release).

Judges also described reviewing and deferring to the PSA-DMF recommendations. Notably, although both P2 and P4 had stated that they primarily focus on the case details, they also described regularly considering and deferring to the recommendations. P2 described deferring to the PSA-DMF model suggestion for cases at the extremes of the risk score values.

> "The more polar the recommendation: if I'm on a one, then I'm going to cut to the one and I'll be good with the one. And the same with the six, I'll go, "Yeah, the

guy's got a ton of risk factors." Now, if I'm in the 3, 4, 5, then I'm going to start going back and looking at the individual risk factors." – P2

This quote describes behavior that is consistent with the findings of a prior experimental study regarding how laypeople respond to PRAIs (Green and Chen 2021).

P4 described following the DMF recommendation in the interest of consistency.

"I would tend to default where there's not a huge difference [between my and the PSA's recommendations]. [...] For purposes of consistency, I would tend to default with what the matrix recommends." – P4

Similar to the prior quotation, this comment suggests that judges selectively defer to the PSA-DMF, based on the observed risk score values and their own internal assessments.

Finally, one judge noted that they were explicitly asked to follow the DMF release recommendation during implementation training to facilitate a social science evaluation of the PSA-DMF's efficacy.

"So, when we deal with PSAs, we've been, because it's kind of a—for lack of a better term—a pilot program, they've asked us to go with the recommendations of the PSA. So when I'm making a decision, often even if I think that the recommendation [...] is a lot to ask someone to do [...] I've been asked to defer to the PSA. [...] All of the training that I've gone to kind of says "less is more" on a lot of these types of cases, but we've been asked to go with the PSA because they've been trying to assess how effective those are." – P3

This passage conflicts with the claim by Arnold Ventures and policymakers that judges are not required/expected to follow DMF recommendations (Arnold Ventures 2019a; New Jersey Courts 2017)—more detail in Section 6.

In sum, judges neither completely disregarded nor completely followed the PSA-DMF recommendations. Instead, judges reasoned about the case at hand and compare their own assessment with the assessment provided by the model. Several judges who explicitly claimed to focus on the case details over the model's report generally also described considering the model's recommendations when making final decisions. Because judges (claim to) use the model's recommendation differently depending on the particular details of each individual case, it is particularly important to understand whether and how model explanations alter this decision-making process.

5.2 How Judges Use Counterfactual Explanations

After judges made initial decisions, we presented them with CFEs and asked whether they would like to update their decisions.

The judges initially treated counterfactuals as factuals. When presented with the CFEs, judges almost always treated the counterfactuals as real changes to the defendant's profile, rather than as hypothetical changes, and updated their decisions based on this information.

"I was assuming in answering that, that I was supposed to assume that one of these risk factors had changed." – P4

[When viewing a counterfactual in which a defendant had one FTA instead of two:] "So now I'm going down and I see that he's only got one failure to appear. Is that what I'm looking at now?" – P8

As a result, judges were particularly bothered when the counterfactual included a pending charge, because they wanted to factor the alleged crime into their decision-making process. It was difficult for judges to reconcile that, because the CFE applies to the model, *any* pending charge—whether serious or not—would alter the model's recommendation in the exact same way. This phenomenon follows the behavior described in Section 5.1, where judges focused on the charge details when making pretrial release decisions.

[Given a CFE in which an additional failure to appear leads to a 'B' and an additional pending charge leads to a 'C' from an 'A':] "It's more so the pending charge rather than the failure to appear. I mean it's a factor, but it's not as big of a factor as the pending charge." – P5

[Looking at the CFE, consisting of 5 options leading to more restrictive recommendations:] "If there is a pending charge at the time of the offense, that could certainly change [my decision]" – P6

"If he has a pending charge at the time of the offense, well, I would want to know what the actual pending charge was." – P8

Interpreting counterfactuals as factuals confused judges about how counterfactuals could generate changes in both directions. Given that judges interpreted counterfactuals as real changes to the profile of the defendant in question, they were particularly concerned by sets of counterfactuals involving changes to the defendant's profile that led to both more and less restrictive release recommendations. Of course, if these changes were factual, then they could not be simultaneously true.

"Ok I'm not clear—I'm looking at one set that says 'less restrictive' and one that says 'more restrictive', and I'm sorry for being dense here, but I don't have sufficient questions for me to answer each one of them under the 'less' model and each one of them under the 'more' model, and my answers would be different depending on whether the less restrictive recommendations are in place or the more restrictive recommendations are in place. So I'm not sure what to do—I'm at an impasse. [...] It appears to me that you're offering me two choices and one set of answers. [...] I don't see that I have a way to answer both the less restrictive and the more restrictive." – P1

"All right, now I'm a little confused by the instrument. [...] So if I got two red boxes, I got one going less and I got one going more, which release conditions... [...] and my box on the left, option one drops him to a two, [... but the question] doesn't tell me which one he is. Is he the less one or is he the more one?" – P2

[After receiving the CFE:] "Am I supposed to take into account both options 1 and 2? [...] Are these two taken collectively? Both option 1 and option 2 have changed now?" – P3

"So when I'm answering this question, am I answering it for the red box on the left or the red box on the right?" – P8

Once judges understood that counterfactuals were hypothetical, they generally ignored them. When judges seemed to be misinterpreting CFEs as factual, we emphasized that the counterfactuals were hypotheticals designed to illustrate the properties of the algorithmic PRAI, not an actual revision to the defendant's profile. At this point, the judges stopped considering the counterfactuals altogether.

[After realizing that counterfactuals are not revisions to the defendant's profile:] "I would still choose the same. No change." – P3

"Oh! Oh! [...]. I was assuming in answering that, that I was supposed to assume that one of these risk factors had changed. [...] I don't know what I put in response to the other ones, but these should be identical. [referring to the questions before and after the explanation]." – P4

"No [I would not change my decision], because nothing has changed. It's a little awkward thought process. [...] What you are saying is: these didn't happen. If they didn't happen, [...] how will I change my recommendation or my decision? That's where I'm losing the logic. [...] If it didn't happen, why would I change my decision?" – P6

[After realizing the defendant's profile has not changed:] "No, then if he's the same person he was to begin with, then I would still just release him [...] because my decision is mostly based on the charge itself and the lack of failures to appear or anything like that, so I think I would leave it the same." – P7

"If the person's different, that's definitely relevant to me. So if these things, hypothetically, if in fact these things are true, I might change my recommendation. [...] [After confirming that the CFEs are merely hypothetical:] Well, I'm going to do the same thing [as before viewing the counterfactuals]. I'd just even be more firmly convinced I was right." – P8

One judge described circumstances in which they could envision changing their responses based on counterfactuals. This judge said that they might revisit their original decision, placing less trust in the PSA-DMF, if a hypothetical change to the defendant's profile caused a large change to the DMF recommendation (e.g., from release option B, "1 phone contact per month, court reminder notifications" to E, "Electronic monitoring/home detention, etc."):

"If that one factor is enough to move it that far, then that would cause me concern that the original recommendation is not restrictive enough." – P4

[When asked: If the counterfactual showed a several step increase or decrease in the recommendation, would your decision change?] "Yeah. I would probably change my original condition. [...] It would give me some pause about the underlying math, if you will. It would cause me to wonder what's going on with the PSA. Why is it this sensitive in making that kind of a change? And I would probably ask questions about the PSA, and what's going on, and why is it doing this? And use it as an example to ask somebody like you to explain to me why it's doing that." – P4

The same judge noted one more circumstance in which the CFE might change their decision. For each judge, we had noted that hypothetical changes might allow them to disregard risk factors that they think are not important. The judge in question responded that they generally would not use the counterfactuals to disregard risk factors. But they might use the "sensitive" risk factors as guides to gathering additional information (e.g., extenuating circumstances explaining a prior failure to appear) during an in-person pretrial release hearing.

"I think if [...] in a particular case I didn't think [a factor] was particularly important [I might use the counterfactual]. I mean, if it was failures to appear older than two years, but I had information in front of me that suggested, two years ago, this was the position that the defendant was in. I now know, two years later, this is the position the defendant is in, and I'm persuaded that this person has sort of turned their life around. They've been through treatment. They've kind of gotten their act together, and that maybe I therefore shouldn't treat this more than two year old stuff as seriously, then knowing that that adjustment could be made, if you take that factor out, then that adjustment could be made. I could see using it more on an individual case, as opposed to using it generally across cases, which means it really would be used, not in the context of a pretrial release decision made on paper, but a pretrial release decision made in court with the lawyers able to give me more and more meaningful information about someone's background." – P4

6 Discussion

Implications of how judges interacted with CFEs of the PSA-DMF. In order to better understand whether explanations can improve human-algorithm collaborations in the context of pretrial risk assessments, we conducted think-aloud trials to study how judges use PRAIs and CFEs when making pretrial release decisions. We hoped to learn about whether providing more insight into a PRAI's operations through CFEs, i.e., how sensitive the model is to marginal changes in defendant characteristics, would cause judges to alter their trust in the model, seek additional information, or otherwise adjust their decisions. For instance, one might expect reliance on a PRAI to decrease as the model's sensitivity to input factor differences increases.

We found that CFEs did not alter the judges' decisions. All eight judges mistakenly treated the counterfactuals as factuals rather than as indicators of model sensitivity. In high-stakes scenarios, there could be significant risks associated with decision-makers misinterpreting counterfactuals as factuals. For instance, if a judge detains a defendant because they misinterpret a *counterfactual* prior conviction as an *actual* prior conviction, the PRAI (with a CFE) would likely violate the defendant's due process rights. Furthermore, once judges did understand what the CFEs meant, they ignored them (rather than used them to understand the model's sensitivity). This behavior was consistent across all eight judges interviewed, and suggests that achieving the potential benefits from CFEs will not be straightforward.

More broadly, our findings suggest that there are many challenges associated with improving human-algorithm collaborations using explanations. In response to evidence about the errors and biases of high-stakes algorithms, many policymakers have promoted explanations to help people understand algorithmic advice and use it responsibly (Green 2022).[1] However, both laypeople and practitioners have been shown to interact with explanations in ways that diverge from what engineers and policymakers expect. Our findings build on prior work demonstrating that, despite their intuitive appeal, algorithmic explanations do not always improve people's ability to interpret and act on algorithmic recommendations (e.g., Bansal et al. (2021); Green and Chen (2019b); Jacobs et al. (2021); Lai and Tan (2019)).

One important implication of our work is that explanations should be developed to align with the types of reasoning that the intended users employ in practice. Although judges struggled to understand the CFEs we provided, they are in fact familiar with counterfactual reasoning. As law students, judges are trained to think counterfactually. For example, first-year students are taught to reason through a negligence lawsuit in part by analyzing "but-for" causation. This analysis asks whether the plaintiff would still have been harmed had the defendant acted more reasonably. Many employment discrimination cases similarly depend on whether a manager would have made the same adverse employment decision if the plaintiff's race or gender were different.

However, CFEs do not align with the counterfactual reasoning that judges typically employ. Legal counterfactuals help lawyers make direct comparisons between what a litigant did and hypothetical alternative actions. In contrast, the CFE presented changes to defendant profiles—aspects of their criminal history that were outside of the defendant's control by the time they were arrested for the present charge(s). These explanations were intended to show how the model's output would have changed following specific edits to a profile that the defendant (by definition) could not have changed. Because counterfactual thinking in judicial decision-making is usually limited to alternative actions the defendant could have taken before being sued or criminally charged, judges first misinterpreted our counterfactuals as *real* changes to a defendant's choices and thus their profiles. When reminded that the counterfactuals could not have been real because defendant profiles are always fixed, the judges ignored them. In other words, judges are likely to dismiss alternative information about the PRAI because nothing has changed or could have changed about the defendant's past, and the release decision does not depend on counterfactual thinking about the defendant's current behavior.

This gap between the counterfactual logic familiar to judges and counterfactual logic presented by our explanations could explain why the judges struggled to understand the CFEs and ignored the explanations once they understood them. In order for CFEs to improve human decision-making, it will likely be necessary to align the explanations with the counterfactual logic already employed by practitioners operating in that context. This may also require much more comprehensive training for practitioners.

Implications of how judges interacted with the PSA (without CFEs). We also gathered information about how judges perceive/interact with the PSA-DMF without CFEs.Our interviews suggest that judges view the PSA-DMF as having a relatively small influence on their decision-making process. Judges stated that they primarily focus on the case/charge information and the attributes of defendants, i.e., the risk factors. They also expressed doubts about the utility of the PSA algorithm. These interviews align with ethnographic studies showing that judges and other criminal justice practitioners resist the use of algorithms, in part due to a sense of professional autonomy and fear of deskilling (e.g., Brayne and Christin (2020)).

However, these claims by judges may not tell the full story. First, several judges also stated that they defer to the PSA-DMF in certain circumstances; even judges who claimed to focus on the case details said that they review the PSA scores before making a final decision and often defer to them. Second, prior work shows people tend to underestimate the effects that algorithmic advice has on their behavior (Green and Chen 2019a,b, 2021). Thus, even if judges state that they do not closely follow the PSA-DMF, the instrument could still influence their behavior in ways that they do not recognize (e.g., via anchoring/automation biases).

One particularly surprising comment was P3's statement that they have been asked to defer to the PSA when making decisions. As noted above, this statement contradicts the claim by Arnold Ventures and policymakers that judges are not required or expected to follow DMF recommendations (Arnold Ventures 2019a; New Jersey Courts 2017). This statement also does not reflect the exact instructions given at the implementation training. Rather, judges were told that, although they retain final discretion, if they never followed the model's recommendations, it would be impossible to evaluate the model's efficacy.[2] In other words, this judge has internalized a stronger directive to follow the DMF recommendations than the training explicitly provided. This

[1] We note that the term explanation can carry many meanings. Here, we are referring to local explanations about a specific decision. Calls for transparency about the overall workings of an algorithm before and during deployment fall into a different (and better trod) category.

[2] Direct observation from one of this paper's authors, who attended the training.

(mis)interpretation complicates the claims that judges retain full discretion when making pretrial decisions. It is therefore necessary to evaluate not just the explicit instructions provided by pretrial risk assessment developers, but also how judges interpret and act on those instructions. Even if judges do not view an instruction to follow the PSA-DMF recommendation as binding, such an instruction could generate automation bias, which would diminish the role of their independent judgment in decision-making.

Limitations. Our study was small, with a study population of just eight judges. However, we observed a high level of consistency in how all eight judges responded to CFEs. Given the current paucity of user studies for CFEs, the detailed reactions described in this paper provide useful insights into how experienced decision-makers respond to counterfactuals. Moreover, these findings can inform future work involving larger-scale user studies with CFEs.

Secondly, we explored just one approach to presenting CFEs. Judges' initial confusion about the meaning of CFEs might have been a function of our UX design. Similarly, the judges' reluctance to use the hypothetical information provided by the counterfactuals could have resulted from insufficient training at the start of the survey. After observing the responses of the first four judges, we attempted to address this possibility by adding an interactive presentation that provided more information about what the CFEs mean and how they could be used. Following this training, all four Round 2 judges were able to summarize what the explanations meant. However, this additional training did not alter the judges' decisions-making. It is possible that more in-depth training or alternative presentations of the counterfactual information would have led to significantly different behavior. An important direction for future work will be to explore whether alternative methods for presenting CFEs would lead to different outcomes.

Our study was also limited to specific jurisdictions where the PSA-DMF has been in use. We chose judges from such locations intentionally, so that our participants would already be familiar with the basic PSA-DMF before adding CFEs. However, the judges' prior experience with the PSA-DMF could have affected how they responded to the CFEs. Because these judges had previously used the PSA-DMF without explanations, they might have been confused about how to incorporate the CFEs into their reasoning. We attempted to account for this potential issue by providing more instructions to participants in Round 2, which helped judges understand the CFEs during training more quickly than in Round 1. However, once the judges encountered CFEs about specific cases and understood that they were hypothetical, they largely ignored them (as they did in Round 1). The fact that judges exhibited the same behaviors even after additional training suggests that there might be a fundamental conflict between CFEs and judicial reasoning, rather than just a lack of understanding of the CFEs.

Finally, it is important to recognize that judicial think-aloud responses might not provide complete insight into their behaviors in practice. Judges might behave differently if they were making decisions with real-world stakes. Moreover, what judges tell us about their reasoning may not reflect their actual reasoning. People are notoriously bad at accurately describing their cognitive processing (Nisbett and Wilson 1977)—a phenomenon that has also been observed in human collaborations with algorithms (Green and Chen 2019a,b, 2021). In the case of judges, there may be particularly strong motivations to downplay the influence of PRAIs out of fear that these tools diminish their autonomy and professional status. Any disconnect between how judges think they use and how they actually use algorithms could have significant consequences. If judges mistakenly believe that a PRAI does not influence their behavior, they may be less likely to scrutinize the advice that it provides. It is therefore essential to pursue mixed-methods research on how people use algorithms and explanations, cross-referencing qualitative insights (both think-aloud and ethnographic) with quantitative analyses of human behavior.

Future Work. Our results suggest several directions for future work. First and foremost, future work should evaluate CFEs with a larger sample of judges and a wider sample of decision-makers. Such work could investigate whether judges uniquely find CFEs unintuitive and unhelpful. The think-aloud results provided here should be further interrogated through large-scale experimental studies that evaluate whether and how CFEs alter human decision-making. Such studies should explore methods for teaching users what CFEs represent; our results suggest that even highly educated and experienced decision-makers can struggle to understand what the explanations mean. In order to thoroughly study how people use CFEs, we must first be able to present CFEs in ways that are comprehensible and actionable.

Second, future work should consider alternative methods for presenting CFEs. Our study explored just one particular design for displaying CFEs. As P4 suggests, decision-makers might find CFEs more helpful if they can (interactively) query a system for information about specific changes rather than simply receive information containing a long list of changes. For instance, if a judge could ask the model how its recommendation would change in light of a specific alteration to a defendant's profile—perhaps in light of extenuating circumstances for a prior failure to appear—judges might find the CFEs more intuitive and more helpful.

Conclusion. We conducted think-aloud trials to study how judges use PRAIs and CFEs when making pretrial release decisions. We found that judges initially mistook the counterfactuals as factual changes to defendants. Once the judges understood the CFEs, they ignored them. This behavior was consistent across all eight judges interviewed. Our findings suggest that using (at least these kinds of) explanations to improve human and AI collaboration is not straightforward, highlighting the importance of evaluating XAI systems with their intended users. Here, we suspect the gap between the counterfactual logic judges encounter in their legal training and the logic of CFEs presented a challenge, explaining the judges' surprising yet consistent reactions. Future work should therefore explicitly consider the types of reasoning employed by the intended users when developing XAI systems (including the instructions and UX).

Acknowledgements

YY acknowledges support from IBM Research, the Miami Foundation, and NSF IIS-1750358. We thank Derek Bambauer, Zana Buçinca, Isaac Lage, Todd Proebsting, and Andrew Woods for helpful feedback and discussions.

References

Adler, P.; Falk, C.; Friedler, S. A.; Nix, T.; Rybeck, G.; Scheidegger, C.; Smith, B.; and Venkatasubramanian, S. 2018. Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54(1): 95–122.

Albright, A. 2019. If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions. *The John M. Olin Center for Law, Economics, and Business Fellows' Discussion Paper Series*, 85.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias. *ProPublica*.

Arnold Ventures. 2019a. Public Safety Assessment FAQs ("PSA 101"). https://craftmediabucket.s3.amazonaws.com/uploads/Public-Safety-Assessment-101_190319_140124.pdf. Accessed: 2022-09-07.

Arnold Ventures. 2019b. Statement of Principles on Pretrial Justice and Use of Pretrial Risk Assessment. https://craftmediabucket.s3.amazonaws.com/uploads/Arnold-Ventures-Statement-of-Principles-on-Pretrial-Justice.pdf. Accessed: 2022-09-07.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PloS one*, 10(7): e0130140–e0130140.

Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M. T.; and Weld, D. 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*, 1–16.

Barabas, C.; Virza, M.; Dinakar, K.; Ito, J.; and Zittrain, J. 2018. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Conference on Fairness, Accountability and Transparency*, 62–76. PMLR.

Binns, R.; Van Kleek, M.; Veale, M.; Lyngs, U.; Zhao, J.; and Shadbolt, N. 2018. *'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions*, 1–14. New York, NY, USA: Association for Computing Machinery. ISBN 9781450356206.

Brayne, S.; and Christin, A. 2020. Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts. *Social Problems*.

Byrne, R. M. 2007. *The rational imagination: How people create alternatives to reality*. MIT press.

Byrne, R. M. 2016. Counterfactual thought. *Annual review of psychology*, 67: 135–157.

Byrne, R. M. 2019. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *IJCAI*, 6276–6282.

Cowgill, B. 2018. The impact of algorithms on judicial discretion: Evidence from regression discontinuities. *Unpublished Manuscript, Columbia Business School*.

Dodge, J.; Liao, Q. V.; Zhang, Y.; Bellamy, R. K. E.; and Dugan, C. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, 275–285. New York, NY, USA: Association for Computing Machinery. ISBN 9781450362726.

Engstrom, D. F.; Ho, D. E.; Sharkey, C. M.; and Cuéllar, M.-F. 2020. Government by algorithm: Artificial intelligence in federal administrative agencies. *NYU School of Law, Public Law Research Paper*, (20-54).

Gaddis, S. M. 2017. How black are Lakisha and Jamal? Racial perceptions from names used in correspondence audit studies. *Sociological Science*, 4: 469–489.

Goyal, Y.; Wu, Z.; Ernst, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Counterfactual Visual Explanations. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2376–2384. PMLR.

Green, B. 2020. The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 594–606. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.

Green, B. 2022. The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45: 105681.

Green, B.; and Chen, Y. 2019a. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 90–99. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Green, B.; and Chen, Y. 2019b. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Green, B.; and Chen, Y. 2021. Algorithmic Risk Assessments Can Alter Human Decision-Making Processes in High-Stakes Government Contexts. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

Grgić-Hlača, N.; Engel, C.; and Gummadi, K. P. 2019. Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 178:1–178:25.

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

Hannah-Moffat, K.; Maurutto, P.; and Turnbull, S. 2009. Negotiated Risk: Actuarial Illusions and Discretion in Probation. *Canadian Journal of Law & Society/La Revue Canadienne Droit et Société*, 24(3): 391–409.

High-Level Expert Group on AI. 2019. Ethics Guidelines for Trustworthy AI. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf. Accessed: 2022-09-07.

Hilton, D. J.; McClure, J. L.; and Slugoski, B. R. 2005. The course of events. *The psychology of counterfactual thinking*, 9: 44–73.

Human Rights Watch, ed. 2017. *"Not in it for justice": how California's pretrial detention and bail system unfairly punishes poor people*. New York, N.Y.: Human Rights Watch. ISBN 978-1-62313-460-0. OCLC: ocn989883656.

Jacobs, M.; Pradier, M. F.; McCoy, T. H.; Perlis, R. H.; Doshi-Velez, F.; and Gajos, K. Z. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry*, 11(1): 1–9.

Julia Angwin, J. L.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*.

Kahneman, D.; and Tversky, A. 1981. The simulation heuristic. Technical report, Stanford Univ CA Dept of Psychology.

Keane, M. T.; Kenny, E. M.; Delaney, E.; and Smyth, B. 2021. If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques. In Zhou, Z.-H., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 4466–4474. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Koepke, J. L.; and Robinson, D. G. 2018. Danger ahead: Risk assessment and the future of bail reform. *Wash. L. Rev.*, 93: 1725.

Lai, V.; and Tan, C. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*, 29–38.

Lewis, D. 2013. *Counterfactuals*. John Wiley & Sons.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Martens, D.; and Provost, F. 2014. Explaining data-driven document classifications. *MIS quarterly*, 38(1): 73–100.

McCloy, R.; and Byrne, R. M. 2000. Counterfactual thinking about controllable events. *Memory & Cognition*, 28(6): 1071–1078.

Miller, T.; Howe, P.; and Sonenberg, L. 2017. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*.

Mueller, S. T.; Hoffman, R. R.; Clancey, W.; Emrey, A.; and Klein, G. 2019. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *arXiv preprint arXiv:1902.01876*.

New Jersey Court System. 2018. Pretrial Release Recommendation Decision Making Framework. https://www.njcourts.gov/courts/assets/criminal/decmakframwork.pdf. Accessed: 2022-09-07.

New Jersey Courts. 2017. One Year Criminal Justice Reform Report to the Governor and the Legislature. https://www.njcourts.gov/courts/assets/criminal/2017cjrannual.pdf. Accessed: 2022-09-07.

Nisbett, R. E.; and Wilson, T. D. 1977. Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review*, 84(3): 231.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Roese, N. J. 1997. Counterfactual thinking. *Psychological bulletin*, 121(1): 133.

Sheriff's Justice Institute. 2016. Central Bond Court Report. https://issues.chicagoreader.com/pdf/20161026/Sheriff_s-Justice-Institute-Central-Bond-Court-Study-070616.pdf. Accessed: 2022-09-07.

Steinhart, D. 2006. A practice guide to juvenile detention reform: Juvenile detention risk assessment. *Baltimore, MD: The Annie E. Casey Foundation*.

Stepin, I.; Alonso, J. M.; Catala, A.; and Pereira-Fariña, M. 2021. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9: 11974–12001.

Stevenson, M. T. 2018. Assessing Risk Assessment in Action. *Minnesota Law Review*, 103: 303–384.

Stevenson, M. T.; and Doleac, J. L. 2021. Algorithmic risk assessment in the hands of humans. *Available at SSRN 3489440*.

UK Information Commissioner's Office. 2020. Guidance on the AI Auditing framework.

Verma, S.; Dickerson, J.; and Hines, K. 2020. Counterfactual Explanations for Machine Learning: A Review. *arXiv preprint arXiv:2010.10596*.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2018. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 31: 841–887.

Woodward, J. 2005. *Making things happen: A theory of causal explanation*. Oxford University Press.

Word, D. L.; Coleman, C. D.; Nunziata, R.; and Kominski, R. 2008. Demographic aspects of surnames from census 2000. *Unpublished manuscript, Retrieved from* https://www2.census.gov/topics/genealogy/2000surnames/surnames.pdf.