Original Research Article

Law Enforcement and Legal Professionals' Trust in Algorithms

Journal of Law and Empirical Analysis
2025, Vol. 0(0) 1–20
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2755323X251325594
journals.sagepub.com/home/lex
Sage

Ryan Kennedy[1], Lydia Tiede[2], Amanda Austin[2], and Kenzy Ismael[2]

Abstract

AI algorithms are increasingly influencing decision-making in criminal justice, including tasks such as predicting recidivism and identifying suspects by their facial features. The increasing reliance on machine-assisted legal decision-making impacts the rights of criminal defendants, the work of law enforcement agents, the legal strategies taken by attorneys, the decisions made by judges, and the public's trust in courts. As such, it is crucial to understand how the use of AI is perceived by the professionals who interact with algorithms. The analysis explores the connection between law enforcement and legal professionals' stated and behavioral trust. Results from three rigorous survey experiments suggest that law enforcement and legal professionals express skepticism about algorithms but demonstrate a willingness to integrate their recommendations into their own decisions and, thus, do not exhibit "algorithm aversion." These findings suggest that there could be a tendency towards increased reliance on machine-assisted legal decision-making despite concerns about the impact of AI on the rights of criminal defendants.

Keywords

artificial intelligence, criminal justice, sentencing, trust, algorithms, experiments

The introduction of artificial intelligence (AI) into the American criminal justice system is transforming policing, prosecution, defense, and judging (Baker et al., 2021).[1] Law enforcement is increasingly using AI to name suspects, forecast crime, expose fraud, prevent traffic accidents, determine cause of death, analyze DNA, detect gunshots, recognize weapons, decipher license plates, identify potential victims of crime, predict recidivism, anticipate crowd behavior, and uncover criminal networks (Rigano, 2019). Legal professionals, such as judges, lawyers, and paralegals, use AI to predict litigation outcomes (Katz et al., 2017), determine bail (Angwin, Larson, et al., 2016; Dressel & Farid, 2018), conduct eDiscovery (Dixon, 2020; Grossman & Cormack, 2011), synthesize federal court data (Adler et al., 2023), write and review contracts (Rich, 2018; Yamane, 2020), analyze case history law (Rigano, 2019), and summarize evidence (Rigano, 2019). For several of these tasks, AI is becoming the standard practice. At least 20 federal law enforcement agencies—and local law enforcement agencies across 49 states—have access to facial recognition technology (Mac et al., 2021). This technology is regularly used by almost half of all police departments in cities with at least a million residents (Goodison & Brooks, 2023). Additionally, 11 states and 178 counties in other states use algorithms to assess criminal risk (Movement Alliance Project & MediaJustice, 2024). Many new applications of AI to the criminal justice system are on the horizon. For example, the U.S. government is funding research to use AI to recommend specific lengths of federal sentences (Brown, Pezewski, & Straub, 2021).

Law enforcement and legal professionals have noted the many potential benefits of AI, such as the more consistent application of the law across cases (Ward, 2019). This professional community, however, is also aware of the potential for misuse of algorithms, which may undermine the due process rights of defendants, an especially vulnerable subpopulation. In particular, legal professionals have levied substantial criticism at the *State v. Loomis* decision that gave

[1]Department of Political Science, Ohio State University, Columbus, OH, USA
[2]Department of Political Science, University of Houston, Houston, TX, USA

Corresponding Author:
Lydia Tiede, Department of Political Science, University of Houston, 3551 Cullen Boulevard, Room 447, Houston, TX 77204, USA.
Email: lbtiede@uh.edu

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

developers of recidivism risk algorithms the ability to maintain the secrecy of their code at the expense of criminal defendants' due process rights to understand the methodology behind the algorithm's recommendations (Freeman, 2016; Liu et al., 2019). The controversy surrounding the *Loomis* decision suggests that the law enforcement and legal community may have conflicting opinions about the increasing integration of AI throughout the criminal justice system. Concerns expressed by law enforcement and legal professionals about the impact of AI on vulnerable populations—particularly in the absence of safeguards and standards for transparency—raise questions about whether these professionals are willing to incorporate it into their own workflows (Greenstein, 2022).

Although there is substantial literature on the ethics, efficacy, and fairness of using AI to assist law enforcement and legal professionals (e.g., Alikhademi et al., 2022; Angwin et al., 2016; Barabas, 2020; Dempsey et al., 2023; Farayola et al., 2023; Miron et al., 2021; Rademacher, 2020; Surden, 2021; Taylor, 2023; Završnik, 2020) and some work assessing how much trust the public places in algorithms to make important decisions in the criminal justice arena (Kennedy et al., 2022; Ozer et al., 2024; Zhang & Dafoe, 2020), this literature fails to specifically explore the perspectives of the professionals most impacted by the integration of AI, specifically subpopulations such as those involved in the criminal justice system (Gerlich, 2023, 2024; Mousavi Baigi et al., 2023; Zhang, 2023). Because the professionals who increasingly wield AI in the criminal justice system differ considerably from the general population in terms of education, demographic backgrounds, training, and work experience, we follow previous research that has emphasized the importance of specifically examining their unique perspectives on the use of AI in their field (Gerlich, 2023, 2024; Hannah-Moffat, 2015; Stevenson & Doleac, 2019; Simmler et al., 2023). We build upon this literature by empirically examining to what extent does the legal and law enforcement community trust machine-assisted advice and incorporate it into their decision-making in comparison to advice received from human sources? This question is an urgent one as the more criminal justice stakeholders trust and are willing to incorporate algorithms' advice into their professional decision-making, the more AI will be used—and possibly misused—in practice.

Across three survey experiments, we assess the trust law enforcement and legal professionals, such as paralegals, lawyers, and judges, place in algorithms within the increasingly common contexts of forecasting recidivism and facial recognition matching for issuance of warrants, as well as within the prospective context of recommending sentence lengths. We choose the contexts of facial recognition and recidivism algorithms because they involve the AI technology most likely to be encountered by law enforcement and legal professionals. We also investigate the context of AI-aided sentence recommendations so we can explore how law enforcement and legal professionals perceive algorithms that they are unlikely to have yet encountered but are likely to confront in the future. With this information, we not only can assess trust in widely used AI applications in criminal justice, but also predict how law enforcement and legal professionals might respond to the introduction of new AI technology in their field.

Our work incorporates the best practice methodologies from research on trust. Our experiments analyze not only *stated trust* in AI, but also *behavioral trust*, which indicates whether respondents have enough confidence to use advice from AI in specific factual situations. Additionally, unlike previous research on this subpopulation, our experiments include a benchmark for trustworthiness. Specifically, we compare the trust that these professionals place in algorithms to that placed in common human sources of advice.

Our findings indicate that although our respondents do not explicitly express trust in AI, they are nonetheless willing to use AI in professional contexts. The first experiment finds that respondents incorporate advice from an algorithm more than from their peers or an experienced judge when forecasting future offenses—a key component in bail, parole, and probation decisions. Our second experiment finds that our respondents rely on advice from algorithms more than anonymous informants and eyewitnesses when issuing warrants based on facial recognition technology. Our third experiment finds that respondents use advice from an algorithm to the same degree as advice from a judge when making decisions related to sentencing after a conviction, although they are most likely to rely on advice from prosecutors or an experienced probation officer.

In sum, despite law enforcement and legal professionals' statements expressing misgivings about AI to assist in key investigatory and legal decisions, our respondents are willing to use the advice of such systems in concrete situations as much as or more than human sources of advice in every scenario in which algorithms are already commonplace. These findings suggest that—in spite of several high-profile controversies about whether the use of AI in the criminal justice process negatively impacts defendants' due process rights (Michaels, 2024; Sachoulidou, 2023)—algorithms are likely to be increasingly incorporated into the workplace as law enforcement and legal professionals demonstrate a willingness to adopt their recommendations.

This research contributes to the scholarly field on the role of trust in algorithms in professional settings by addressing several key gaps in the academic literature. It introduces novel scenarios involving a relatively unexplored population, particularly within the judicial community, where AI's role in decision-making is critically relevant. Previous studies have been criticized for not focusing on populations most impacted by AI implementation, which this research rectifies. Additionally, the study differentiates between individuals' stated level of trust in AI and actual willingness to incorporate AI into their decision-making, revealing a complex relationship

between what people say and how they act. Despite long-standing, vocal concerns about the use of algorithms in sentencing, this study finds that people continue to rely on algorithms for decision-making, regardless of their stated trust.

1. Legal Professionals' Trust in Algorithmic Decision-Making

Previous studies have reached divergent conclusions about trust in algorithms among the general public as well as specific groups (Choung et al., 2022; Mousavi Baigi et al., 2023). Some scholars note individuals' algorithm aversion, in which individuals trust algorithms less than humans (e.g., Dawes, 1979; Dietvorst et al., 2015, 2018, p. 114; Gogoll & Uhl, 2018, pp. 97–103). Other scholars question the existence of this algorithm aversion (Lee, 2018; Logg, 2016). Some research suggests that individuals often prefer decisions made by machines over humans (Gerlich, 2023, 2024; Kennedy et al., 2022) or at least favor hybrid decision-making, in which both types of decision-makers are used (e.g., Fry, 2018, p. 71; Ramakrishnan, et al., 2014; Scharre, 2018). Moreover, some studies show that individuals are unlikely to go against algorithmic recommendations (Kennedy et al., 2022). However, individuals often prefer their own judgment over that of any other decision-maker, whether a human or machine (White & Eiser, 2010). In some cases, however, individuals may actually prefer algorithms over humans due to the perceived impartiality of algorithms in making decisions (Gerlich, 2024). Additionally, some scholars have concerns about using algorithms—not due to their effectiveness—but due to the fact that they may perpetuate racial bias (Angwin, Varner, & Tobin, 2016) and their use in certain areas, such as criminal sentencing, may raise ethical concerns (Taylor, 2023). None of these findings, however, necessarily carry over to the subpopulation of law enforcement and legal professions, as they differ markedly from the general public.

Understanding the extent this subpopulation trusts and relies upon the recommendations generated by AI algorithms is critical for several reasons. The use of AI in law enforcement and legal communities is growing rapidly (Maslej et al., 2023, p. 291), with ideas for novel applications emerging at a swift pace. This technology can have major impacts on the lives of vulnerable and marginalized segments of society (Angwin, Varner, & Tobin, 2016). For example, previously unsolvable crimes may be brought to justice (Brewster, 2023), criminal activity may be identified and stopped before crimes are committed (McDaniel & Pease, 2021), cases may be processed more quickly (Chen et al., 2022), criminal defendants may spend more or less time behind bars (Ryberg & Roberts, 2022), and demographic disparities in criminal justice administration can be increased or decreased as a result of technological advances in AI (Ho et al., 2023). Additionally, the incorporation of AI may impact the way law enforcement and legal professionals conduct their work. If these professionals distrust AI, then they may be disinclined to use it and instead turn to human decision-makers who also may prove to be inconsistent, biased, or prone to mistakes. On the other hand, law enforcement and legal professionals' overconfidence in legal advice generated by AI may result in its unquestioned use without accountability mechanisms in place. Moreover, trust in AI may change the tenor of the relationships at the heart of the criminal justice system and implicate important ethical issues (see Taylor, 2023, for example, for ethical issues related to use of AI in sentencing). The use of AI in cases may impact the level of trust that law enforcement and legal actors have in one another or the duties, responsibilities, and advice that lawyers and other legal actors owe to defendants (Goodman, 2019). Finally, understanding trust in AI's use in law enforcement and legal settings is crucial as significant distrust in legal tools and processes may undermine the legitimacy of law enforcement and legal systems (Tyler, 2006, p. 115).

Legal and law enforcement professionals' trust in AI depends—in part—on whether algorithms in criminal justice have worked or are just (Surden, 2021), but the findings on their impact are mixed. Some authors argue that the incorporation of AI has worsened pre-existing injustices in the criminal justice system, such as racial bias and the treatment of vulnerable populations (Angwin, Varner, & Tobin, 2016; Dressel & Farid, 2018; but see Fry, 2018, pp. 66–68; Okidegbe, 2021, 2022). Other scholars suggest that algorithms may outperform judges who are inconsistent in their decisions (Austin & Williams, 1977; Dhami & Ayton, 2001; Kleinberg et al., 2017; Lin et al., 2020). Still, others find that AI has "little impact on either the judge's decision and subsequent arrestee behaviour" (Imai et al., 2023). The inconsistency of these findings makes it difficult to predict whether AI engenders trust among law enforcement and legal professionals.

The literature also fails to make clear whether any trust these professionals place in AI in the abstract carries over to the specific contexts in which legal professionals use AI in practice. For example, a National Judicial College survey finds that 65% of the 369 judges surveyed believe that AI is a "useful" tool for dealing with bias in decision-making, but "should never completely replace a judge's discretion" (Firth, 2020). Additionally, the Thomas Reuters Institute (2023) surveyed law firms and found that 82% of respondents think ChatGPT and generative AI "can be readily applied to legal work," but only 51% indicate it "should" be used. Although these surveys provide a useful overview of attitudes towards AI among the legal community in the abstract, law enforcement and legal professionals may not necessarily hold consistent positions about AI's abilities both in general and in specific instances.

Moreover, existing scholarly work concerning law enforcement and legal professionals' trust in AI tends to

concentrate more on their stated trust in AI rather than whether they will actually use it in their professional capacities. The most common understanding of trust in the social sciences comes from McKnight & Norman (1996), who argue that trust is the "willingness to be vulnerable based on positive expectations about the actions of others." The traditional way to measure people's trust in a person, institution or system has been to ask them a Likert-style question. However, just because individuals express trust or lack thereof in AI does not necessarily mean they will or will not rely on it in practice. Indeed, research has shown surprisingly little correlation between stated trust in computational systems in Likert-style questions and people's propensity to actually use those systems or behavioral trust (Elhamdadi et al., 2022; Xiong et al., 2019). Thus, understanding both the stated and the behavioral trust, as measured in actual decisions, that actors place in AI is critical, as the two forms of trust may not necessarily align.

Additionally, much of this literature fails to provide a counterfactual to the use of AI to make recommendations regarding matters of justice. Studies indicate that survey participants often discount advice from AI due to factors such as reliance on their own knowledge or a general aversion to algorithms (Jussupow et al., 2020; Kennedy et al., 2022). However, these findings do not accurately represent the counterfactual scenario to using AI advice, which in this context would be better illustrated by considering advice from other human sources. In other words, few scholars probe the *relative* trust of law enforcement and legal professionals in advice received from AI and from human actors through comparative causal analysis (but see Gerlich, 2023, 2024). Since law enforcement agents and legal professionals regularly receive advice from human sources, such as their peers and legal experts, and also receive feedback from the public in the form of elections, advice from human sources is a natural point of comparison to advice received from AI. Although studies of relative behavioral trust in algorithms exist, they examine the general public rather than our subpopulation of interest (Kennedy et al., 2022; Logg, 2016).

We posit that law enforcement and legal professionals may be more willing to incorporate advice given by AI than a human source into their decision-making regarding criminal cases for several reasons. First, law enforcement and legal professionals may trust AI in the criminal justice sphere precisely because it is already integrated into many of their workplace procedures. Any first-hand experience using AI in practice may predispose these professionals to trust the technology (Maslej et al., 2023, p. 291). With greater exposure to algorithms in routine criminal justice cases, such as those that determine the likelihood of recidivism, these professionals may appreciate that algorithms work most of the time or at least as well as human advisors. For example, in technical systems with which people have extended interaction, such as autopilot systems or factory robots, trust in such systems has been found to build over time, even resulting in overtrust (Wagner et al., 2018). Further, endorsement by the courts in cases such as *Loomis* "communicate to judges that the tools are in fact reliable" (Harvard Law Review, 2017, p. 1536).

Second, law enforcement and legal professionals may trust AI more than the general public trusts AI because they have—on average—higher levels of education or experience. Some public survey results about AI show that levels of trust vary by education (Logg et al., 2018; but see Lee & Baykal, 2017; Thurman et al., 2019). Thus, among a population that has a higher average level of education than the public, we might expect to see greater trust in AI.

Third, law enforcement and legal professionals may prefer advice from AI to humans due to sheer self-interest or self-preservation, a point made by Zhang (2023). For lawyers who perceive that AI helps or hurts their clients or their professional practice, they may express more or less trust in its use. However, this level of trust may vary by profession, as those in law enforcement and paralegals may see algorithms as undermining their traditional role as a key source of advice. Thus, the use of algorithms may exhibit heterogeneous effects depending on the category of professional.

Finally, law enforcement and legal professionals may prefer advice from algorithms to humans because such advice is viewed as more objective (Gerlich, 2024). For example, critiques of judicial decision-making argue that judges decide cases based not only on the law and facts but also on their own biases (Lynch & Omori, 2018; Wooldredge et al., 2005), including political preferences (Segal & Spaeth, 2002). Moreover, the objectivity of judges' decisions has been questioned by the correlation between judges' favorable rulings and how close it is to lunchtime or even whether their local football team won a recent game (Chatziathanasiou, 2022).

We investigate our hypothesis with three survey experiments that probe how the law enforcement and legal community weigh advice received from different AI and human sources within the context of specific criminal justice scenarios. Our survey experiments test whether these professionals are more likely to use advice from algorithms over advice from human sources. We follow each experiment with a text analysis of an open-ended question posed to our respondents that explores their stated trust in the technology.[2]

2. Sample

For our sample, Surveys USA solicited 1,019 respondents with experience in the law enforcement and legal fields who participated in the survey experiment between October 14 and November 12, 2022. There are five categories of respondents: law enforcement, paralegal, lawyer/law student, judge, and other legal jobs (See, Supplemental Materials, SM1, for a list of these "other" jobs). Since some of our respondents have acted in multiple roles during their careers, these professions sum to over 100. Although all our

respondents have experience in law enforcement and the law, it was not possible to narrow this sample further to only legal professionals who, for example, exclusively practice criminal justice law.[3] We did not screen for experience with algorithms in their job because interviews conducted prior to the survey suggested that most legal professionals did not recognize often-used tools in their practice as algorithms or AI. Nonetheless, given that few people who work with these systems in the future will have specialized expertise in how algorithms work, we believe this provides a useful baseline for understanding legal professionals' reactions.

Table 1 shows that our sample is more highly educated, male, and white than the U.S. general population, although this skew aligns with the demographics in the legal field more generally. For example, 82% of our respondents identified as white and 44.8% identify as female, which is consistent with demographic breakdowns for legal professionals in the U.S. According to 2020 data from the American Bar Association, about 86% of all lawyers identify as "non-Hispanic white" and women make up about 38% of all lawyers (ABA, 2020).

3. Study 1: Trust in Risk Evaluation

The use of algorithms for estimating the risk of defendant re-offense for decisions on setting bail, pre-trial detention, probation, and sentencing is one of the most common and controversial applications in criminal justice (Fry, 2018, p. 65). Judges, probation officers, parole boards, and parole officers must not only make decisions about the severity of the crime but also the defendant's probability of recidivism if released. Defense attorneys, prosecutors, and—to some extent—the paralegals who work with them use recidivism forecasting to inform their legal strategy for or against specific bail amounts and sentence lengths. Correctional facilities can even consider these forecasts when making decisions about allocating resources and implementing crime prevention strategies within jails and prisons. The decisions made by these professionals inherently involve a prediction task about the likelihood of future recidivism. As a result, these professionals usually rely on algorithmic assessments about defendants' likelihood of reoffending found in presentencing investigation reports (PSIs). This use of algorithms to predict recidivism has defied many legal challenges, as shown by *Loomis* and other cases (Michaels, 2024). Because of the ubiquity and sensitivity of these algorithmic risk evaluations, they have become a touchstone of many of the debates about algorithm ethics, bias, and efficacy.

Despite the prevalence of risk assessments based on algorithms, we know little about whether and to what extent law enforcement and legal professionals trust them enough to incorporate them into their practice. Therefore, in our first study, we explore trust in such algorithms to forecast criminal recidivism.

Table 1. Summary Demographic Characteristics of Sample.

Respondent Type	*N*	Mean	St. Dev.	Min	Max
Law enforcement	1,019	0.294	0.456	0	1
Paralegal	1,019	0.307	0.462	0	1
Lawyer	1,019	0.424	0.494	0	1
Judge	1,019	0.028	0.166	0	1
Other legal job	1,019	0.217	0.412	0	1
Female	1,019	0.448	0.498	0	1
White	1,019	0.820	0.384	0	1
Education	1,019	4.913	1.209	2	6
Party affiliation	986	3.866	1.974	1	7

3.1. Experimental Design

Our design builds on studies by Dressel and Farid (2018) and Kennedy et al. (2022). Respondents view the profile of a criminal defendant and predict how likely the defendant will be to commit another felony within two years. Profiles came from a database of 2013–2014 pretrial defendant records from Broward County, Florida, which contains information on the demographics of the defendant, whether they committed a felony in the past, and their risk scores from COMPAS, which assigns scores of 1–10, with 1 indicating the least risk of re-offense and 10 the highest (Dressel & Farid, 2018). This public record information was compiled by *ProPublica* (Angwin et al., 2016). Although in some cases decision-makers would have far more detail about each defendant, the amount of information presented in these profiles is the minimum amount that is commonly provided in recidivism forecasting (for survey item wording examples see Supplemental Materials, SM2).[4] The design of the profiles is similar to those used by Kennedy et al. (2022), in that we select 20 random profiles with COMPAS re-offense risk scores between 3 and 8 to avoid obvious cases and pair them down by removing extremely unusual profiles or obscure violations in consultation with a field expert. From there, we select 8 profiles at random. Using these profiles allows direct comparisons between the responses of the general public in previous research and the responses of the legal community in our research. The defendant profiles read:

> The defendant is a [male/female] aged [age]. They have been charged with [offense]. This crime is a [misdemeanor/felony]. They have been convicted of [number of prior convictions] prior crimes. They have [number of juvenile felony charges] juvenile felony charges and [number of juvenile misdemeanor charges] juvenile misdemeanor charges on their record.

Similar to the COMPAS algorithm, respondents are provided with the probability that defendants would commit another crime in the next two years. Dressel and Farid (2018) used this wording with online respondents (on MTurk) to achieve similar accuracy to COMPAS.

After submitting forecasts, respondents are shown the following paragraph, with the source of advice randomized:

> A(n) [previous survey of 286 people like yourself/judge with more than 10 years of experience adjudicating criminal charges of both adults and minors/algorithm[5] developed by computer scientists and criminal justice researchers, based on a statistical analysis of thousands of past defendant records], [on average] rated the defendant as [advice] likely to commit another felony crime within the next two years. Previously, you forecasted that the defendant was [Respondents' previous forecast] likely to commit another felony crime within the next two years. If you would like to update your forecast, you can do so now. If not, just enter the same numbers as you entered previously.

These three sources of advice draw from distinct theories. The public survey represents the "wisdom of the crowd" (Suroweicki, 2005). Law enforcement and legal professionals usually receive advice from the public indirectly, via elections and other forms of public feedback. The experienced judge represents the role of expertise (Ozer, 2023). Law enforcement and legal professionals receive expert advice directly via training. The algorithm represents AI as a supplement to human decision-making (Fry, 2018, p. 71; Ramakrishnan et al., 2014; Scharre, 2018). Although discussion of AI as a replacement for human decision-making abounds (e.g., Chen et al., 2022), in practice AI generally serves as an aid to human judgment rather than a replacement. The wording for the three sources of advice was primarily driven by what we considered a minimal amount of information decision-makers would likely have at their disposal. We attempted to avoid any extraneous wording that might indicate past accuracy. Treatments differ slightly in terms of their length, but there is little evidence length independently influences experimental responses (Taber et al., 2009). The forecasts were the same for all three sources of advice.

For our dependent variable, we calculate two quantities from the responses (Gino & Moore, 2007; Kennedy et al., 2022; Logg, 2016; Poursabzi-Sangdeh et al., 2021; Yaniv, 2004). *Absolute distance to advice* is calculated as $|a_i - u_{2i}|$, where a_i is the advice regarding the probability the defendant will recidivate and u_i is the final respondent forecast after seeing the advice. *Weight of advice* is calculated as $|u_{2i} - u_{1i}| \div |a_i - u_{1i}|$, where a_i is the advice related to the defendant's likelihood of recidivism, u_{1i} is the respondent's initial forecast, and u_{2i} is the respondent's final forecast. Weight of advice ranges from 0 to 1, where zero means that the legal professional respondent gave no weight to the advice and 1 means the respondent completely aligned their forecast with the advice. Both of these measures capture how the respondent's behavior, i.e., their forecasts, change based on treatment assignment.

The results are pooled across defendant profiles and respondents. Since we expect heterogeneity on both levels, we utilize a multilevel problem of the form:

$$y_i \sim N\left(\alpha_i + \beta(Treatment), \sigma^2\right)$$

$$\alpha_i = \alpha_0 + \alpha_{profile[i]} + \alpha_{respondent[i]}$$

where y_i is either the distance from the advice or weight of advice, α_i is the individual intercept, and β is the estimate of the average treatment effect (ATE). The individual intercept is a function of the overall intercept, α_0, plus the random intercept for the defendant profile and for the respondent. Because there is no pure control condition, i.e., where no advice is present, we use the treatment where advice is given by an experienced judge as the baseline. This baseline is chosen because judges are the standard decision-makers in the context of bail or sentencing decisions, and are tasked, by law, with making these decisions on the basis of such risk (Barabas, 2020).

In addition to the ATE, we also check how differences in respondents' backgrounds might affect their responses. As noted by Zhang (2023), it is important that studies related to algorithm trust examine subpopulations of respondents who are particularly affected by the use of algorithms. As such, we subset our sample by professional category. We exclude judges as a category from this analysis because our sample—much like the general population of law enforcement and legal professionals—has a small percentage of judges and, thus, insufficient power to estimate moderation. We then estimate the average treatment moderation effects (ATME) using parallel within-treatment regression analysis (Bansak, 2021). This approach has the advantage of causal interpretation, which is absent from conditional average treatment effects (CATEs) estimated using treatment-by-covariate interactions. In the multilevel situation, this has the form:

$$\begin{aligned} y_i &\sim N\left(\alpha_{0ikl} + \beta_{0Si} + x_{iyi}, \sigma^2\right) \quad \forall i : T_i = 0 \\ y_j &\sim N\left(\alpha_{1jkl} + \beta_{1Sj} + x_{jyj}, \sigma^2\right) \quad \forall j : T_j = 1 \end{aligned}$$

where S is the moderator of interest, i is the case where the value of the treatment is 0, and j is the case where the treatment is 1. Responses in each model are still modeled as nested within respondent (k) and scenario (l). The ATME, δ_{PR}, is estimated as $\delta_{PR} = \beta_1 - \beta_0$ and the variance of the ATME is $Var(\delta_{PR}) \sim Var(\beta_0) + Var(\beta_1)$.

To better understand our survey results, we also ask respondents an open-ended question following the experiment. While the experiment details *whether* respondents alter their responses in response to various sources of advice, it cannot explain *why* respondents may or may not respond to treatments. Open-ended items help uncover respondents' rationales and thought processes, as well as shed light on the distinction between respondents' stated trust in using algorithms and their behavioral trust. To evaluate these open-ended responses, we use both structural topic modeling (STM) and natural language processing (NLP) sentiment analysis. STM uses unsupervised machine learning to identify topics within preprocessed text and to assign words to each topic (for text preprocessing and model details see Supplemental Materials, SM3). STM then represents the text

as a vector of proportions showing what fraction of words belong to each topic and incorporates covariates of interest into the prior distributions for document-topic proportions and topic-word distributions (Roberts et al., 2014). We also use Natural Language Processing (NLP) to uncover the sentiment expressed by respondents. To do so, we employ three state-of-the-art methods: (1) Stanza NLP sentiment analyzer, which uses a convolutional neural network (CNN) to classify the sentiment of sentences in the responses (Qi et al., 2020); (2) Valence Aware Dictionary and sEntiment Reasoner (VADER), which evaluates words within a n-gram context to determine the valence as well as sentiment (Hutto & Gilbert, 2014); and (3) TextBlob, which uses the sentiment of the words in a sentence to develop a sentence score (Loria, 2018). To enhance the performance and robustness of these methods, we combine them into an ensemble (Zhang & Ma, 2012).

3.2. Findings

Figure 1(a) shows the weight given to advice from all three sources. Participants give significantly more weight to their own judgment than advice from any other source, consistent with prior research (Imai et al., 2023). The vast majority of respondents, 79%, give less than 50% weight to the advice provided in the experiment from algorithms, judges, and a survey of the public (for tabular versions of all results see Supplemental Materials, SM4). This result is also consistent with Kennedy et al.'s (2022) findings from a similar experiment on a national sample, where 72% give less than 50% weight to the advice provided. In particular, legal professionals are much more likely to trust their own judgment than advice from a survey of the public ($p < .001$). More respondents give 50% or more weight to the advice received from the judge than from the survey of the public, showing there is not much behavioral trust in the "wisdom of the crowd."

Of the three sources of advice, respondents gave the most weight to the advice from the algorithm. This finding is illustrated in Figure 1(b) and (d). In Figure 1(b), respondents' final forecasts are six points further away from using the advice when it originates from the survey of the public than when it is from the judge ($p < .001$). Forecasts are about 1 point closer to the advice provided by the algorithm than by the judge, but the difference is not significant ($p > .05$). Kennedy et al.'s (2022) similar experiment on a national sample found a 3.5-point increase in the distance to advice when it came from the survey of the public and a 2-point decrease in the distance when the advice came from an algorithm, and both were statistically significant ($p < .05$). Thus, the impact of advice from judges and algorithms relative to the public survey is greater among this sample of law enforcement and legal professionals than among a national sample. However, there is less discrimination among law enforcement and legal professionals between advice received from the judge and the algorithm than from the survey of the public.

In Figure 1(d), which accounts for the respondents' initial guess in calculating the weight of advice, respondents give about 12% more weight to advice from the judge than from the survey of the public ($p < .001$). Respondents also give 6% more weight to the algorithm than to the judge, and this difference is statistically significant ($p < .01$). By comparison, in Kennedy et al.'s (2022) experiment fielded on a national sample, respondents gave about 7% more weight to the judge's advice than to the survey of the public ($p < .001$) and 6% more weight to the advice from the algorithm than to advice from the judge ($p < .01$). In our survey, we find larger differences between expert and crowd-based sources of advice among the law enforcement and legal community respondents than Kennedy et al. (2022) found among a national sample, although the difference between the weight of advice given by the algorithm and the judge is quite similar in both samples.

In sum, for both the law enforcement and legal professionals, advice from algorithms is preferred over advice from judges, which in turn is preferred over advice from a survey of the public. However, the results indicate greater differences in opinion among law enforcement and legal professional respondents regarding advice received from single experts (judges and algorithms) than advice received from the public survey. Additionally, Figure 1(c) and (e) demonstrate that the type of law enforcement or legal background respondents have makes no significant difference in the results. Thus, our first experiment finds no detectable differences based on particular professional categories.

To explore the relationship between stated trust and behavioral trust and to determine the rationale behind respondents' answers in the experiment, our open-ended question directly following the first experiment asks respondents: "Can you please elaborate on why you did or did not trust the [survey's/judge's/algorithm's] advice?" Our topic analysis of this open-ended question uncovers two topics, of which the highest word probabilities are presented in Figure 2(a).

Topic one is cited more frequently by respondents in the judge's treatment group (Figure 2(b)). It includes not only the term "algorithm" but also "trust." The concurrence of these two terms occurs frequently in the open-ended text questions, as respondents decide whether to trust algorithmic decision-making. One oft-stated comment is that it is hard to trust the advice of an algorithm without a better understanding of its architecture and training. For example, one respondent said, "I never completely trust computer algorithms, What data base? How did they create it?" Another frequently expressed concern is that an algorithm cannot possibly predict the often-irrational behavior of people. This concern was noted by a respondent who indicated that "I would not trust an algorithm to predict irrational human behavior like the decision to commit a crime." The overall results from the structural topic model underscore the experimental findings that respondents generally trust their own experience first and foremost, e.g., "I

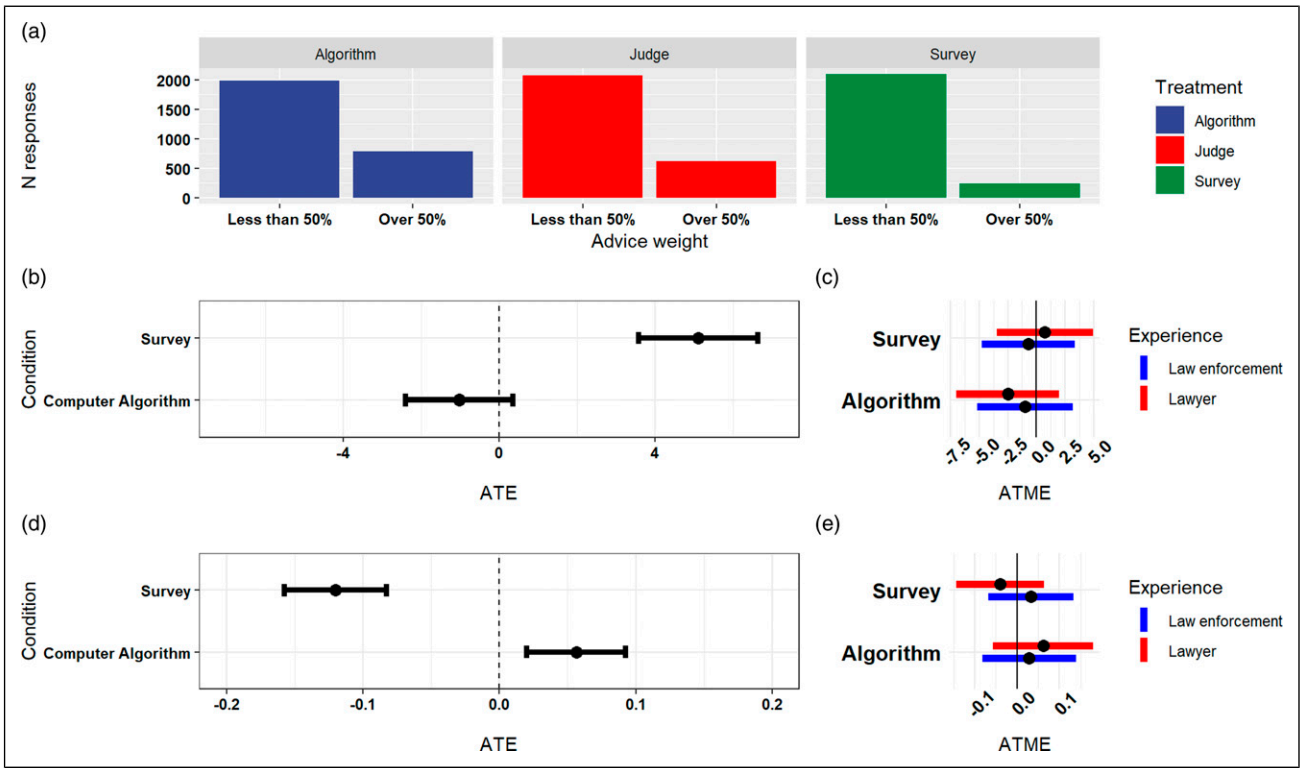

Figure 1. Behavioral trust in risk evaluation sources. Figures show the results across eight scenarios for all 1,019 respondents. Figure 1(a) shows the distribution of the weight of advice for each treatment condition, broken down by whether the respondent in the scenario gave more or less than 50% weight to advice from the various treatment sources. Figure 1(b) shows the results of the multilevel model on absolute distance from advice, with the judge's advice as the baseline source (the dotted zero line). ATE and standard errors are calculated across 1000 simulated draws from the coefficient distribution. Figure 1(c) shows the ATME of law enforcement and legal professional categories, estimated through parallel regression. Figure 1(d) shows the results of the multilevel model for the weight of advice, with the judge as the baseline condition. Figure 1(e) shows the parallel regression estimates of ATME for weight of advice.

trust my experience more than an algorithm." Advice from other sources, including algorithms, is afforded less weight.

For the second topic, "crime" is the word mentioned most frequently, followed by "people." When we incorporate the treatment group to which each respondent is assigned as our covariate of interest, we find that this first topic is more highly cited by respondents who received advice from the survey, than from other sources (Figure 2(b)). Two other terms of note in this first topic are "think" and "see," which speak to respondents' tendency to weigh their own beliefs higher than any other source. Respondents usually attribute their confidence in their own judgment based on the length of time they have worked in the criminal justice system, even indicating the number of years they have worked in their profession. For example, one respondent in the treatment group that received advice from an experienced judge, states: "I have practiced law for almost 40 years… We have a current district judge who was in diapers when I passed the bar. I have more experience and I believe better judgment then [sic] many of them."

Figure 2(c) depicts the sentiment expressed about each source of advice. The results indicate that respondents express slightly more positive sentiment toward advice received from an algorithm than from either a survey of the public or a judge ($\chi^2 = 16.095$; $p = .041$). Although respondents' opinions about algorithms run the gamut from completely trusting to fully rejecting the algorithm's predictions about human behavior, most legal professional respondents express moderate skepticism about the algorithm's recommendations, generally trusting their own judgment above that of the algorithm, while simultaneously exhibiting slightly higher trust in an algorithm than in other sources of advice.

Looking at these responses, the importance of analyzing behavior and stated trust separately becomes apparent. The open-ended responses show that respondents have mixed feelings of trust and mistrust, with the algorithm receiving both the most positive and the most negative assessments of all the sources of advice. In contrast, the experiment demonstrates that despite stated misgivings, respondents are more likely to incorporate advice from an algorithm into their decision-making than from any other source. This disconnect between stated and behavioral trust is very apparent when we home in on the responses of individuals in our sample. For

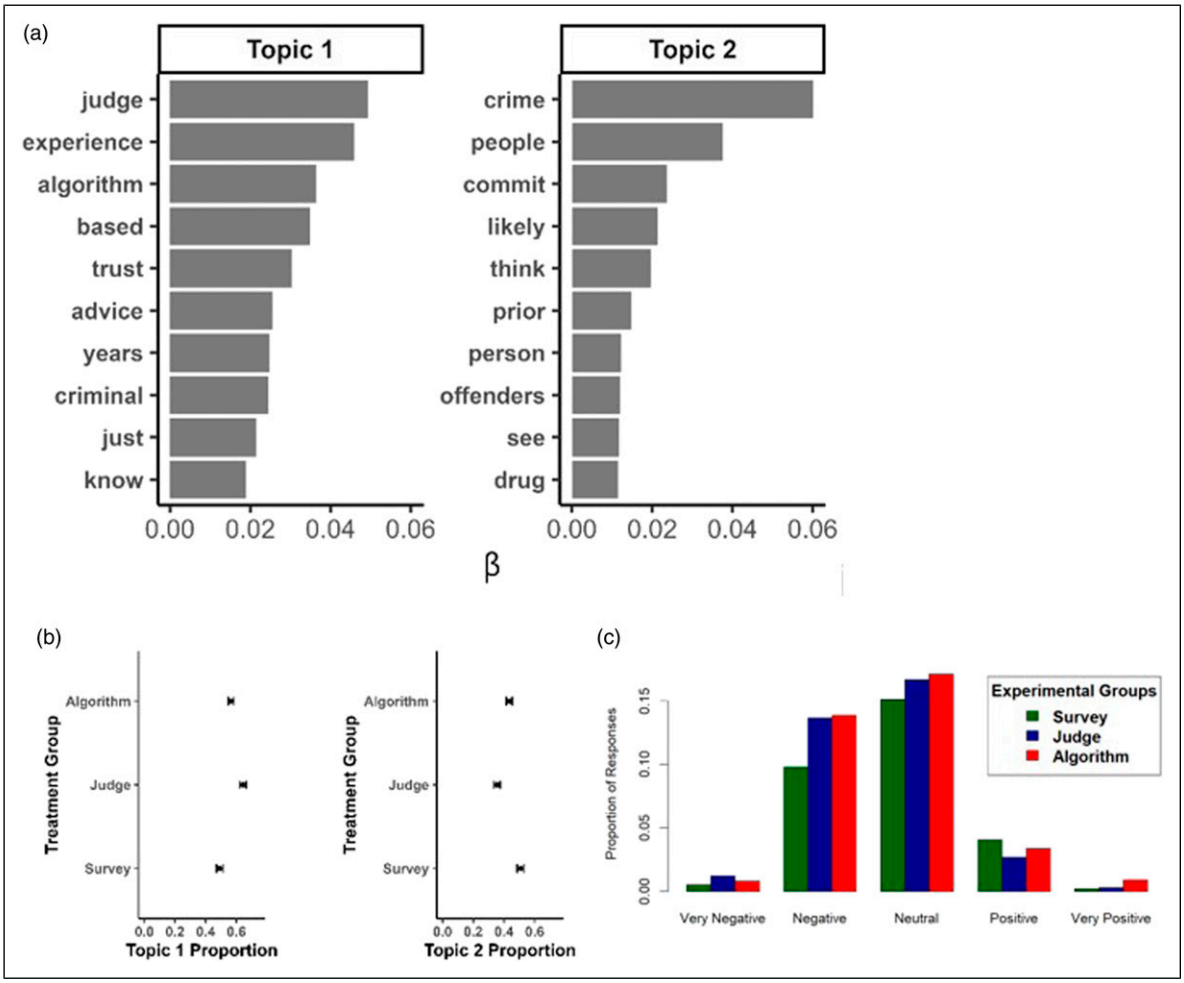

Figure 2. Text analysis of open-ended evaluations of source of advice. Figure 2(a) shows the results of topic modeling. Figure 2(b) shows the differences in topic mentions among treatments or advice sources. Figure 2(c) shows an ensemble sentiment analysis of responses regarding the sources of advice.

example, one respondent stated, "I didn't trust the algorithm's advice because I have no idea what it's calculating," but in the experiment gave 4.5% more weight to the algorithm's recommendations than other advice sources. Another said that they "do not think [the algorithm's forecasts] are right," but gave them 29% more weight than average in their final decisions. Still another said that "Algorithms aren't super reliable for human behaviors," but completely changed their initial predictions in the experiment to match the algorithm's, giving the algorithm's advice 57% more weight than the average respondent in this treatment condition. Conversely, two respondents stated, "I mostly trusted the algorithm" and "I trusted all of it," yet these respondents gave 10% *less* weight to the algorithm's advice than the average respondent in our experiment. Another said that "Algorithms are good" but gave no weight to the algorithm's forecasts whatsoever, saying they went with their "previous experience within the judicial system." Our experimental results and open-ended responses underscore the disconnect between individuals' stated trust in algorithms and their behavioral trust in the use of algorithms in practice, an important—but often ignored—distinction.

Intriguingly, we find that the few respondents who state they *do* trust algorithms do so for a single, major reason: algorithms are accurate. Along this vein, respondents stated: "I trust it because it is logical and clear; " "I was able to trust the algorithm's advice because of the statistical analysis which had been performed;" and "I trusted the algorithms advice more than my own hypotheses, since the algorithm can take in a great amount of data and analyze it for patterns and probabilities." Some respondents noted the accuracy of algorithms in comparison to their own human judgement: "i work around inmates who constantly come to jail for different crimes and from what i have seen is why i did trust the algorithm's advice." Some also believed that algorithms are *more* accurate than their own judgment, making statements such as "My responses, before the algorithm advice, was 100% subjective. The algorithm, I assume, is reasonably

objective. I trust expert opinion more than my own guess." This belief that algorithms are at least as accurate, if not more so, than human judgment forms the basis of trust in their use.

4. Study 2: Trust in Facial Recognition

The use of facial recognition for the identification of criminals is another controversial use of AI (Van Noorden, 2020). Algorithms may produce inaccurate facial recognition, which has led to individuals being falsely accused and convicted and has raised a myriad of worries regarding its misuse (Fry, 2018; Hill, 2020; Responsible AI Collaborative, 2024).

Despite concerns regarding the accuracy of facial recognition algorithms, its use in criminal investigations is growing. Facial recognition software is now available to law enforcement in nearly every U.S. state, and it is estimated that images of about 50% of U.S. adults are in the facial recognition databases used by these programs (Responsible AI Collaborative, 2024). This technology has even expanded to composite use with other AI technologies. A recent investigation revealed that in 2017, detectives in Northern California constructed a face from DNA found at the scene of a crime and then ran the generated face through facial recognition software (Mehotra, 2024). While this instance is the first known use of this composite technology, reporting by *Wired* found that law enforcement representatives were largely open to this novel usage, despite the lack of evidence of its efficacy.

The alternatives to facial recognition, however, are also generally not considered reliable. Humans have highly sophisticated abilities to recognize faces, yet eyewitness identification is often not considered trustworthy (O'Neill Shermer, Rose, & Hoffman, 2011). Having eyewitnesses identify suspects by looking through mugshots is considered problematic since participants feel they need to positively identify at least one of the pictures (Wells, 1988). Anonymous tips, though often sought by police, can be even less reliable and are usually only allowed for evidentiary searches when there is additional corroboration or include significant verifiable details (Bates, 2000).

In this second study, we test law enforcement and legal professionals' trust of facial recognition software versus other types of human identification for supporting the issuance of a warrant to search a suspect's home. Even though law enforcement officials are usually the professionals who directly use facial recognition technology, legal professionals, such as lawyers and paralegals advise their clients about its use, and judges must decide whether to trust the advice of this technology when determining whether to issue a warrant. Because facial recognition technology informs the work of various categories of law enforcement and legal professionals, it is important to understand their behavioral and stated trust in its recommendations.

4.1. Design

To evaluate the degree to which respondents trust facial recognition versus other standard sources to identify a suspect, we provide respondents with the following scenario, randomizing the source of suspect identification:

> Investigators have gone to a court requesting a warrant to search the home where a suspect in a robbery investigation lives. Put yourself in place of the judge, and answer how likely you would be to issue this warrant. [An anonymous informant saw surveillance footage of the crime on TV/An eyewitness to the crime looked through a book of mugshots/A facial recognition program evaluated surveillance footage of the crime] and says they are 75% (probable) sure they recognize the suspect. Investigators believe the suspect resides at the address for which they want the warrant. If you were the judge, how likely would you be to agree to issue the search warrant?

Respondents rate their likelihood of issuing this warrant on a five-point scale from "extremely unlikely" to "extremely likely." This rating is normalized to range from 0 to 1 so that the resulting analysis estimates the proportion change in support on this scale. Here again, instead of measuring trust in terms of stated trust, we measure how their behavior on this question changes with assigned treatment condition.

Since this is a single vignette, estimation of the average treatment effect (ATE) can be done using ordinary least squares (OLS) regression with the treatments entered as dummy variables (Gerber & Green, 2012). We set the anonymous informant as the baseline since it is generally considered the least reliable of the human sources of identification and provides the closest human analogue to a facial recognition program using similar input data. As with Study 1, we test for moderation by the respondents' professional category using parallel regression.

4.2. Results

Regardless of the source of advice, respondents were generally split on issuing a warrant. About 42% said they were unlikely or very unlikely to issue a warrant, while about 48% said they were likely or very likely to issue a warrant. Figure 3 displays the results comparing sources of identification, with Figure 3(a) showing the ATE estimates. The identification by the anonymous informant is the least persuasive source. Both facial recognition and eyewitness identification generate greater support for issuing a warrant ($p < .05$). For both facial recognition and eyewitness identification, there is about 5% higher support for issuing the warrant. The results do not show, however, a significant difference between advice from the eyewitness and the facial recognition technology's identification—the coefficients are nearly identical. This finding suggests that facial recognition technology is as trusted as eyewitness identification—a source that, while problematic, is generally likely to result in a warrant being issued.

Figure 3(b) shows once again that there is not a significant difference in treatment effects among respondents with

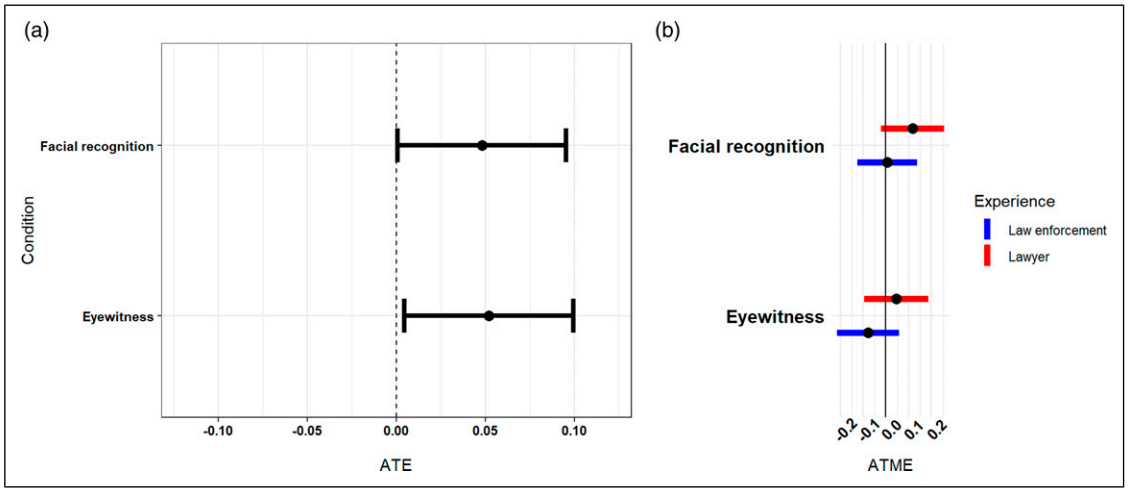

Figure 3. Support for issuing warrants based on different sources of suspect identification. Figures show results from the suspect identification scenario across all 1,019 participants. Figure 3(a) shows the ATE estimated by an OLS regression on support for issuing a warrant with the informant who identified the suspect from a video as the baseline (dotted zero line). Figure 3(b) shows the ATME estimates, done by parallel regression, for different legal backgrounds.

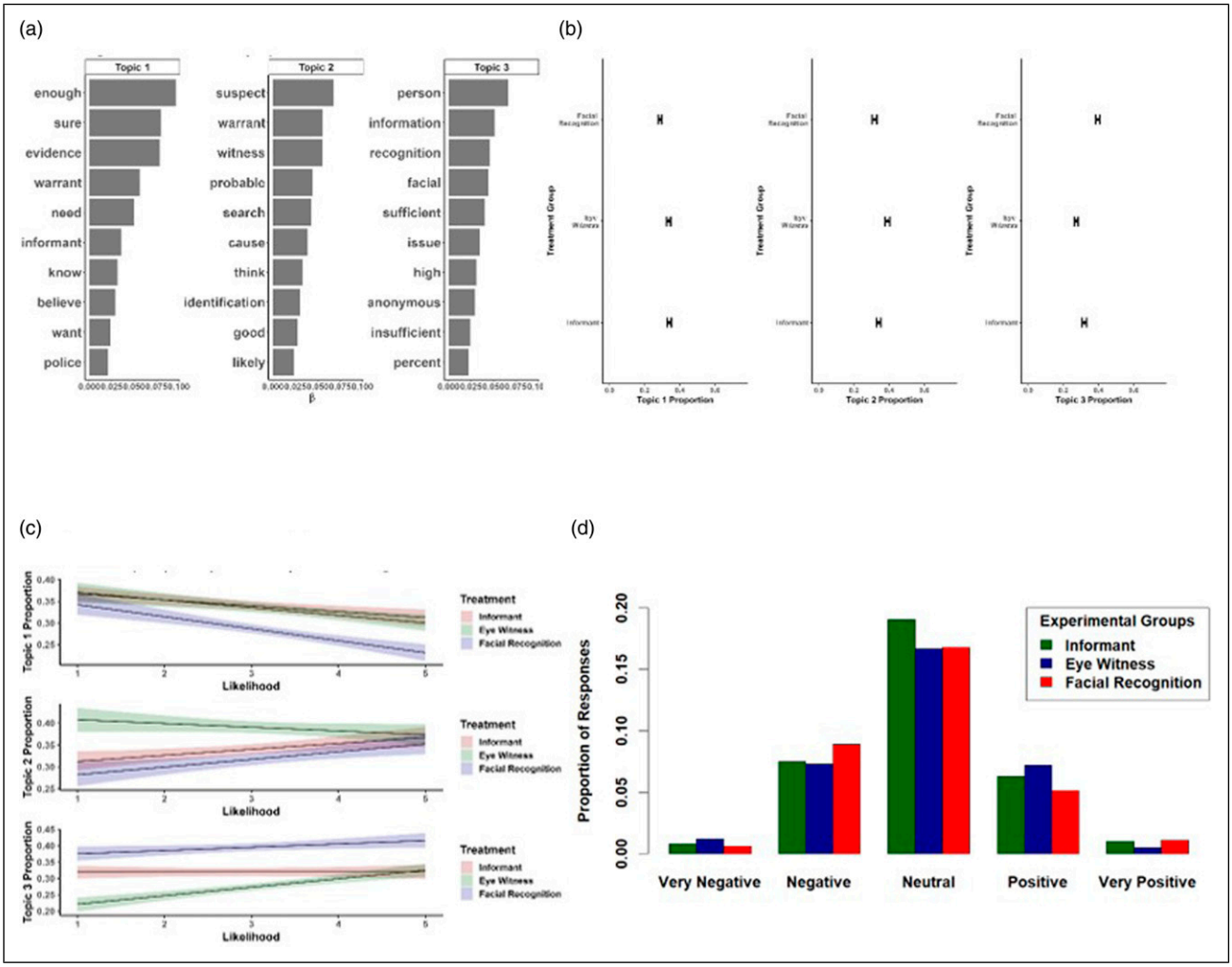

Figure 4. Text analysis of open-ended evaluations of witness identification sources. Figure 4(a) shows the results of topic modeling. Figure 4(b) shows the differences in topic mentions between treatment arms. Figure 4(c) shows the differences in topic mentions by treatment arms and by respondents' likelihood of issuing a search warrant. Figure 4(d) shows an ensemble sentiment rating in responses about the sources of identification.

different professional backgrounds. The closest to a statistically significant difference in effect is between lawyers and paralegals with regards to facial recognition—lawyers seem to be somewhat more likely to use facial recognition identification—but this does not reach conventional levels of significance, so we cannot reject the null hypothesis.

As with Study 1, we use text analysis of an open-ended question following the experiment, asking respondents "Why did you think the [source]'s information was sufficient or insufficient to issue a warrant?" The results of the textual analysis are displayed in Figure 4. The STM identifies three topics. The frequent terms used in Topic 1 reveal doubts about the ability of the source to correctly identify the suspect, the terms in Topic 2 indicate greater confidence, and the terms in Topic 3 demonstrate conflicting opinions or uncertainties regarding the reliability or accuracy of the source's judgment. The STM suggests differences in the qualitative responses given by respondents (Figure 4(a)) and depending upon their treatment group (Figure 4(b)). Respondents in the facial recognition group are slightly less likely to cite Topic 1 words, such as they "need" more "evidence" to be "sure" about whether to issue a warrant and much less likely to cite Topic 2 words, indicating respondents in this group are less likely to believe there is "good" "probable" "cause" to issue a "search" "warrant." In contrast, individuals in the facial recognition group are more likely to use Topic 3 words, which indicate ambiguity, such as "sufficient" and "insufficient" "information." An interaction of these topics with treatment groups and the experiment's dependent variable finds that respondents in the informant and facial recognition group who cite Topic 2 frequently are most likely to issue a warrant (Figure 4(c)). Notably, despite the hemming and hawing over confidence in facial recognition, respondents in this experimental group who cite Topic 3 frequently are highly likely to issue a warrant. This finding suggests that while respondents voice some doubts about the ability of facial recognition software to correctly recognize a suspect, they are still willing to place their trust in the software. The ensemble NLP (Figure 4(d)) analysis bolsters this interpretation, finding no major differences among the sentiments expressed towards the various sources.

Some open-ended responses suggest that a portion of the individuals receiving the facial recognition treatment think that the current software is reliable. Examples of this perspective include: "facial recognition don't lie," "It [facial recognition] is more accurate than a person," "Because facial recognition works," "Facial recognition technology is insanely good now," and facial recognition is "hard to beat." Even respondents who expressed doubts, however, still said they were likely to issue a warrant. Respondents who made statements like "I would want something more," "not until it's efficient," and "I don't trust the data. It can be biased on race" still indicated they were likely to issue the warrant based on facial recognition software—the same as most of the more supportive responses above.

5. Study 3: Trust in Sentencing Guidance

For our final study, we evaluate the degree to which law enforcement and legal professionals trust advice from algorithms regarding the length of sentencing of defendants. This study differs from the evaluation of recidivism risk in Study 1 and the issuance of a warrant in Study 2, as it involves advice on punishment after a final judgment of culpability has been reached, as opposed to forecasting and identification. Placed in the context of Agrawal et al.'s (2018) dichotomy between forecasting and judgment, this experiment places the evaluation of the algorithms in a qualitatively different context. It also differs from our other experimental treatments in that this AI technology has been developed but is not yet deployed in the criminal justice system. Considering the pace at which AI is being introduced into pre-existing processes in the criminal justice system, we believe it is important to explore whether law enforcement and legal professionals are willing to trust new technological innovations that impact their work.

We choose the context of AI recommendations for criminal sentence lengths for several reasons. First, although our respondents are unlikely to have encountered AI sentencing recommendations, it is likely they will do so in the not-too-distant future. The National Science Foundation has funded research on the use of AI to recommend specific lengths of federal criminal sentences (Brown, Pezewski, & Straub, 2021). As the current federal sentencing guidelines already operate as an algorithm, albeit one that does not use AI, the incorporation of AI into sentencing is not farfetched. Indeed, several nations are currently experimenting with algorithms that recommend sentence lengths: Malaysia has piloted the use of an algorithm to recommend sentence lengths for rape and drug convictions (Chandran, 2022) and New Zealand is exploring the use of an algorithm to recommend sentences for assault offenses (Rodger et al., 2023). Second, this context is comparable to that of recidivism predictions and facial recognition in that it is likely to provoke controversy, as any potential biases in the algorithm can conceivably heighten demographic disparities in outcomes among defendants. Third, research finds that a representative sample of the American public trusts sentencing decisions by AI to the same degree that they trust lower-stakes decisions, such as the issuance of consumer refunds (Chen et al., 2022). Assessing whether American law enforcement and legal professionals would similarly trust this novel technology allows us to forecast whether this subpopulation would be likely to adopt it—and other new and potentially controversial uses of AI—into their professional decision-making.

5.1. Design

Like Study 2, this third study uses a single vignette. The respondents are first provided with a scenario of a defendant entering a guilty plea. The respondents are then given state

sentencing guidelines for the crime and are advised by a randomized source that the defendant should receive a lower sentence. The base group is advice from the probation officer since this source of advice is often required for sentencing evaluation in large counties, providing a realistic baseline for evaluation. The crime, sentencing guidelines, and recommendations are structured along actual guidelines and scenarios studied in previous work (Tiede, 2007, 2009). The wording is as follows:

> A defendant has entered a guilty plea to conspiracy to transport a controlled substance (cocaine). The defendant has no past criminal history and the defendant accepted responsibility for this crime. The state's sentencing guidelines suggest a range between 5 and 15 years, although the judge has discretion to go below or above this range. [The prosecutors in the case/A judge with more than 10 years experience who evaluated the case/An algorithm developed by judges and computer scientists evaluated the case and/A probation officer who evaluated the case] suggested that the defendant was low risk for recidivating and recommended a 2-year sentence. How long would you sentence the defendant to prison?

Respondents could choose a sentence between 1 and 20 years. The absolute value between the sentence chosen and the advice provided serves as the dependent variable. As before, this measure is normalized between 0 and 1 so that results can be interpreted as the proportionate change in the distance from advice. As with the other experiments, the goal is to see how the respondent's response to this outcome question changes with the treatment condition, rather than asking directly about trust. Since this study—like Study 2—is a single vignette, analysis of the ATE is calculated using standard OLS regression. Analysis of moderation by legal experience is done using parallel regression. We again follow our experimental results with textual analysis of respondents' answers to an open-ended question.

5.2. Results

Results suggest that views of algorithms relative to other sources of advice differ when it comes to recommending a sentence, even though this task involves evaluation of risk, for which respondents showed a relatively high level of trust in algorithm-generated advice in Studies 1 and 2. Figure 5(a) shows no statistical difference between advice from the probation officer and the prosecutor. For both types of advice, respondents gave nearly identical sentences and were closer to the advice from these sources than from other sources. In contrast, when the judge provides the advice, respondents' sentences are generally higher and further away from the advice. This difference is statistically significant, at least when compared to the probation officer baseline. The distance from the algorithm's advice is also significantly higher than the distance from the probation officer's advice, though it is quite like the sentences given under the judge condition. Figure 5(b) shows that there is not significant moderation by the respondent's legal background.

There are several possible explanations for this result. First, similar to the distinction suggested by Agrawal et al. (2018), legal professional respondents' trust in algorithms for making forecasts and giving probability estimates may not carry over to the making of judgments themselves. Second, the prosecutor and probation officer may have been seen as adversarial sources, such that their recommendation for lighter sentences carries more weight under the "value of biased information" (Calvert, 1985). Third, as the use of AI to recommend sentence lengths is not yet standard in the

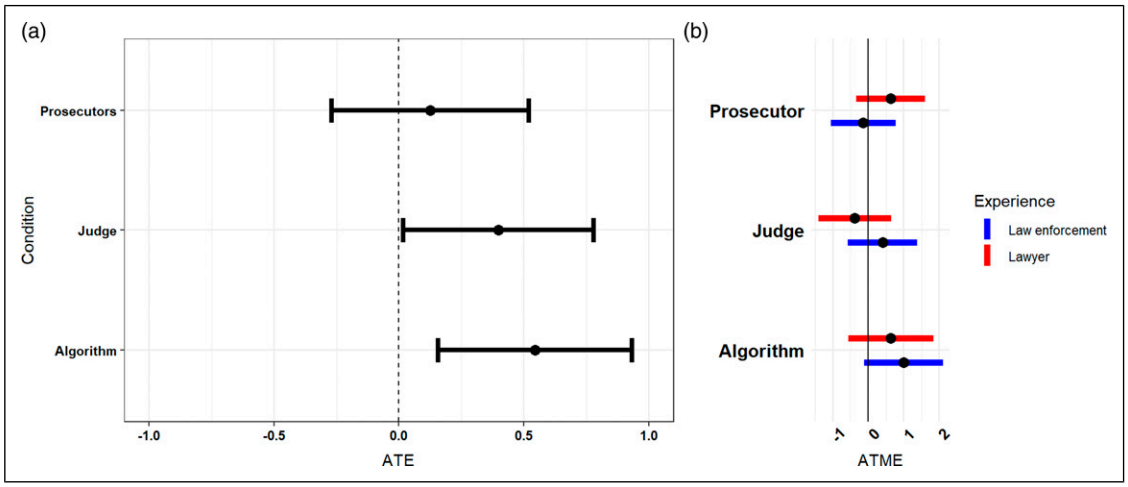

Figure 5. Distance from proposed sentence based on source of advice. Figures show the results from the sentencing scenario across all 1,019 participants, based on the source of the sentencing advice (probation officer, prosecutor, judge, and algorithm). Figure 5(a) shows the ATE of distance from the advice with the probation officer as the baseline category (the dotted line at zero). Figure 5(b) shows the ATME estimates, done by parallel regression, for different legal backgrounds.

American criminal justice system, respondents may remain more skeptical of its accuracy and fairness.

An analysis of responses to an open-ended question also reveals skepticism as to the use of algorithms as a source of advice in sentencing. In this open-ended question (Figure 6), we ask respondents for the basis of their sentencing decisions. Topic modeling (Figure 6(a)) finds that most respondents rely on the facts of the case, defendants' willingness to take "responsibility" for their actions, and the "risk" of recidivism (Topic 1). However, respondents also rely on sentencing "recommendation[s]," especially those made by "probation" officers and "judge[s]." Respondents also refer to the defendant's "history" and "prior" offenses; and sentencing "guidelines," which are all conventional factors of sentencing

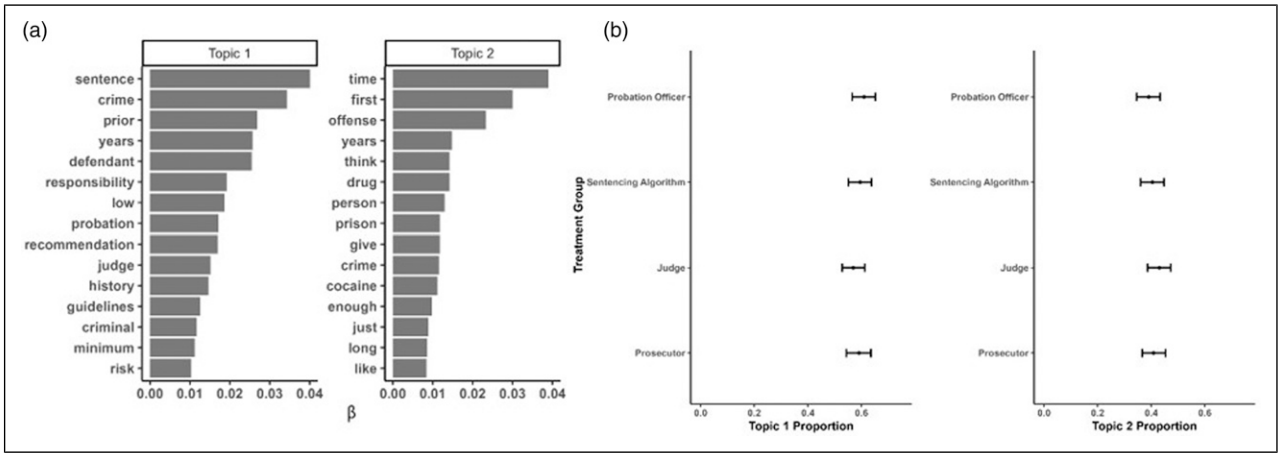

Figure 6. Text analysis of open-ended evaluations of sources of sentencing advice. Figure 6(a) shows the results of topic modeling. Figure 6(b) shows the differences in topic mentions between treatment arms.

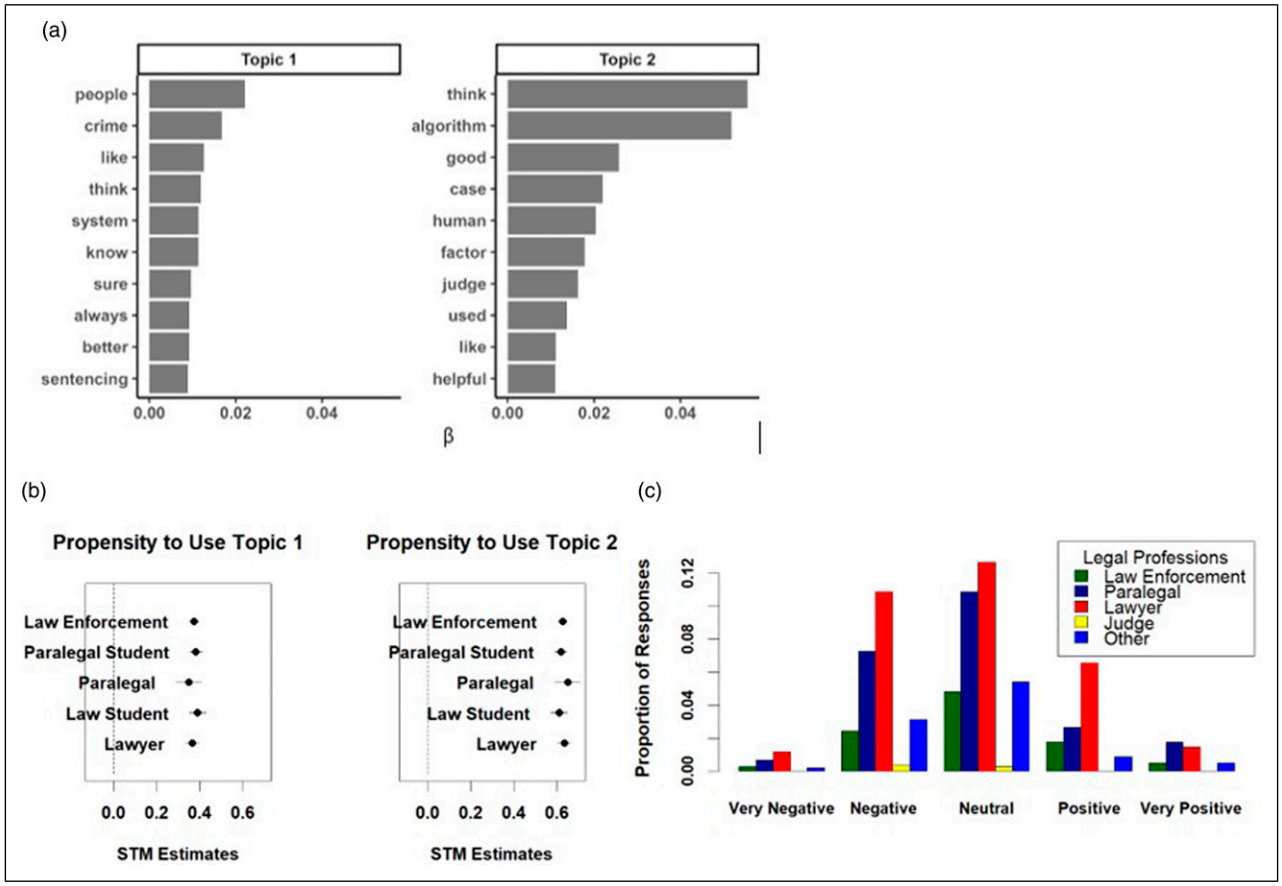

Figure 7. Text analysis of open-ended evaluations of algorithms increasing use in sentencing and bail decisions. Figure 7(a) shows the results of topic modeling. Figure 7(b) shows the differences in topic mentions by varying legal occupations. Figure 7(c) shows an ensemble sentiment rating of respondents' opinions about algorithms.

(Topic 1). Additionally, respondents are likely to vary the "time" they recommend according to the defendant's use of "drug[s]", particularly "cocaine." Respondents rarely mention algorithms in these responses and their references to these topics do not differ significantly between treatment groups (Figure 6(b)).

At the conclusion of our experiments and subsequent open-ended question, we seek to establish a baseline attitude among our sample towards the use of AI in the criminal justice system. To do so, we ask all members of our sample to comment on the trend that "algorithms are increasingly being used to advise on sentencing and bail decisions" (Figure 7). Our results show that many respondents express concern that algorithms are not as "good" as "human[s]" in weighing "factors" and "think[ing]," although they acknowledge that algorithms can be "helpful" (Figure 7(a), Topic 2). We also find that respondents express more negative sentiment towards the involvement of algorithms in the criminal justice system than positive sentiment (Figure 7(b)) with no significant difference among legal professionals of different types (Figure 7(b) and (c)). Respondents argue that algorithms miss the human element, with one respondent stating that the algorithms in sentencing are "just another tool to use but should not replace our human intuition or empathy." This skepticism toward using AI in sentencing mirrors historical criticisms invoked when federal and state sentencing guidelines were first introduced in the United States in the mid-1980s, a more rudimentary type of algorithm itself. Criticisms at that time focused on the ills of replacing human discretion with formulaic sentencing guidelines (Tiede, 2007, 2009). This stated criticism did not stop the incorporation of algorithms in federal sentencing guidelines then, and our results suggest it is not likely to halt the proliferation of AI throughout the criminal justice system now.

6. Conclusion

As hundreds of thousands of criminal cases are currently impacted by the use of algorithms in the United States and as the use of AI in other areas of legal decision-making is likely to expand, research into the attitudes of professionals most likely to deploy this technology is important. The decisions in these cases not only influence individual litigants but also public trust in law enforcement and the courts. Legal decisions involve a tremendous amount of discretion delegated to public officials, such as police, prosecutors, public defenders, and judges. Increasingly, portions of this decision-making are being delegated to machines that employ powerful algorithms. Our results suggest that law enforcement and legal professionals may accept the encroachment of AI on their responsibilities as it becomes more routinely used and endorsed by courts.

Our studies seek to understand not only whether legal professionals *express* trust in algorithms, but whether they are willing to *use* them in practice compared to advice received from other human sources. Unlike other studies, we conduct our survey experiments exclusively with law enforcement and legal professionals, engaging their expertise and experience in evaluating factual cases found within the criminal justice system. Our study is one of the first to compare the willingness of law enforcement and legal professionals to incorporate advice received from algorithms to that of other human sources.

In our first experiment on forecasting recidivism, law enforcement and legal professionals generally value their own experience to make judgments about the factual scenarios provided, confirming past research that finds a bias towards trust in one's own judgment. Additionally, respondents value advice received from an algorithm somewhat over that of a judge, and far more so than advice from a survey of the public. Text analysis of responses to an open-ended survey question reveals that respondents' opinions are grounded in both their own judgment and the facts of the case, including information about the defendants, while expressing mixed sentiment towards algorithm advice.

In the second experiment, legal respondents value facial recognition and eyewitness identification for the justification of a warrant more than an anonymous informant. This result, as with the result of the first experiment, does not reveal algorithm aversion. In fact, our respondents equally value identification from facial recognition as much as from an eyewitness. The open-ended responses provide a more tempered approach to AI deployment, however, as some respondents express concerns that facial recognition may be biased.

In the third experiment, law enforcement and legal professionals value advice from a judge and an algorithm equally, but to a lesser degree than advice from a probation officer or prosecutor. Although respondents do incorporate advice from the algorithm into their sentencing recommendations to the same extent that they incorporate advice from an experienced judge, they simultaneously express concern that algorithms are not sensitive enough to evaluate individual circumstances in decisions concerning the imprisonment of defendants and suggest that sentencing determinations need human input. This same concern about missing the "human element" in sentencing is similar to concerns raised when sentencing guidelines were introduced at the federal and state level in the 1980s and 1990s, which took discretion away from judges. Respondents also, however, expressed appreciation for the objectivity with which an algorithm might calculate criminal sentences as opposed to the subjective and possibly biased calculations by judges. Again, this justification for using algorithms was also present in the debates advocating the introduction of sentencing guidelines in the past.

Overall, these results indicate that in some circumstances where the use of algorithms is already commonplace, law enforcement and legal professionals do not exhibit algorithm aversion and, instead, demonstrate a willingness to rely upon

advice given by algorithms. These professionals do, however, remain skeptical of novel uses for AI in the criminal justice system. The same reasons given for this skepticism, however, are also expressed about more commonly used AI technology that respondents *are* willing to incorporate into their work. Fears about algorithms' accuracy, bias, opacity, and lack of humanity undergird distrust in the use of AI in commonplace scenarios in the criminal justice system, such as predictions of recidivism risk and facial recognition, as well as novel ones, such as sentence length recommendations. As the use of AI becomes more widespread throughout the legal system, we should expect it to be increasingly embraced in practice, if not in principle.

A limitation of our sample is that it is not comprised exclusively of criminal justice professionals who are likely to incorporate AI into their daily workloads, such as detectives and judges. We recommend that future research works within this narrower sampling frame if possible. Our work is also limited by its experimental design. Future work should test the robustness of these findings by comparing them to observational data on the actual incorporation of algorithms into the work of legal and criminal justice professionals. Finally, we encourage researchers to explore this topic in countries outside the U.S. to examine whether these findings hold true across different populations, ensuring that the results are not simply a cultural artifact specific to American society.

Despite the lack of algorithm aversion, legal professionals in our sample place significant weight on their own judgment and demonstrate a nuanced understanding of AI, including awareness of its potential for unethical behavior. These findings suggest that while law enforcement and legal professionals may not blindly adopt AI into their work, they are unlikely to allow it to fully replace human decision-making. If these professionals value both their expertise and the objectivity AI offers, there is potential for AI to enhance legal decision-making—provided concerns about injustice and bias are carefully addressed. However, as this research indicates, even when individuals acknowledge the ethical risks of AI, they may still rely on it for guidance in critical decisions. Therefore, policymakers must recognize that technical and legal guidelines are essential to ensure the ethical and just use of AI in the criminal justice system, even if professionals express awareness of its potential ethical pitfalls.

Acknowledgments

We thank Ken Alper and Rachael Hinkle for useful comments as well as those given during presentations at the American Political Science Annual Meeting in 2023 held in Los Angeles and the International Society of Public Law Annual Conference in 2024 in Madrid.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded, in part, by NSF grant #DASS-2131504, Community Responsive Algorithms for Social Accountability (CRASA). The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies or endorsements, either expressed or implied, of the NSF or the U.S. Government.

Ethical Statement

Ethical Approval

All studies were reviewed by and approved by the Committee for the Protection of Human Subjects (CPHS) at the University of Houston.

ORCID iD

Lydia Tiede https://orcid.org/0000-0003-0892-4649

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Definitions of what counts as artificial intelligence (AI) or not can be very tricky. Russell and Norvig (2010, p. viii) define the field of AI as "the study of agents that receive precepts from the environment and take actions." This is quite broad and includes everything from statistical forecasting models and expert systems to generative AI. The main problem being that there is no agreed-upon definition of what counts as "intelligence." This has led to a joke amongst engineers that, "Whatever we had yesterday is not AI. Whatever we developed today is." As a result, we use the terms "algorithm" and "AI" interchangeably.
2. This research was approved by the Institutional Review Board or Committee for the Protection of Human Subjects (CPHS) at the University of Houston, and the authors certify that the study was performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments.
3. We explored the possibility of conducting more focused surveys on particular subgroups, but found doing so to be prohibitively expensive. We also considered focusing on those who had some interaction with algorithmic decision-making systems. As discussed, there is little indication that these decisions affected the results, given the lack of heterogeneity in responses.
4. In Dressel and Farid (2018), this set was sufficient for workers on Amazon's Mechanical Turk (MTurk) to achieve equivalent accuracy to the COMPAS recidivism risk algorithm.
5. In our vignettes, we default to the term algorithm or computer program due to the difficulties of defining AI as mentioned in footnote 1. Previous studies have used this term as individuals are more able to identify the term algorithms as opposed to AI (Logg, 2016). This delineation also provides consistency with other social science literature.

References

Adler, R. F., Paley, A., Li Zhao, A. L., Pack, H., Servantez, S., Pah, A. R., Hammond, K., & Consortium, S. O. (2023). A user-centered approach to developing an AI system analyzing US federal court data. *Artificial Intelligence and Law*, *31*(3), 547–570. https://doi.org/10.1007/s10506-022-09320-z

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Review Press.

Alikhademi, K., Drobina, E., Prioleau, D., Richardson, B., Purves, D., & Gilbert, J. E. (2022). A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law*, *30*(March), 1–17. https://doi.org/10.1007/s10506-021-09286-4

American Bar Association. (2020). *Profile of the legal profession*. ABA.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Ethics of data and analytics*. Auerbach Publications.

Angwin, J., Varner, M., & Tobin, A. (2016). *Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks*. ProPublica.

Austin, W., & Williams, T. (1977). A survey of judges' responses to simulated legal cases: Research note on sentencing disparity. *Journal of Criminal Law and Criminology*, *68*(2), 306. https://doi.org/10.2307/1142852

Baker, J. E., Hobard, L. N., & Mittelsteadt, M. G. (2021). *AI for judges, a framework*. Professional Report, Center for Security and Emerging Technology.

Bansak, K. (2021). Estimating causal moderation effects with randomized treatments and non-randomized moderators. *Journal of the Royal Statistical Society: Series A*, *184*(1), 65–86. https://doi.org/10.1111/rssa.12614

Barabas, C. (2020). Beyond bias: Re-imagining the terms of 'ethical AI' criminal law. *Georgetown Journal of Law & Modern Critical Race Perspectives*, *12*, 83–111. https://doi.org/10.2139/ssrn.3377921

Bates, E. (2000). Search and seizure - anonymous tips lack sufficient reliability to establish reasonable suspicion for investigatory stop-and-frisks. *Cumberland Law Review*, *31*, 803.

Brewster, T. (2023). *DHS used Clearview AI facial recognition in thousands of child exploitation cold cases*. Forbes.

Brown, L., Pezewski, R., & Jeremy, S. (2021). Determining sentencing recommendations and patentability using a machine learning trained expert system. arXiv preprint arXiv:2108.04088. https://arxiv.org/abs/2108.04088.

Calvert, R. (1985). The value of biased information: A rational choice model of political advice. *The Journal of Politics*, *47*(2), 530–555. https://doi.org/10.2307/2130895

Chandran, R. (2022). *As Malaysia tests AI court sentencing, some lawyers fear for justice*. Reuters.

Chatziathanasiou, K. (2022). Beware the lure of narratives: 'hungry judges' should not motivate the use of 'artificial intelligence' in law. *German Law Journal*, *23*(4), 452–464. https://doi.org/10.1017/glj.2022.32

Chen, B., Stremitzer, A., & Tobia, K. (2022). Having your day in robot court. *Harvard Journal of Law and Technology*, *36*(1), 127–169. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3841534

Choung, H., David, P., & Ross, A. (2022). Trust in AI and its role in the acceptance of AI technologies. *International Journal of Human-Computer Interaction*, *39*(9), 1727–1739. https://doi.org/10.1080/10447318.2022.2050543

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571–582. https://doi.org/10.1037/0003-066x.34.7.571

Dempsey, R. P., Eskander, E. E., & Dubljević, V. (2023). Ethical decision-making in law enforcement: A scoping review. *Psychology International*, *5*(2), 576–601. https://doi.org/10.3390/psych5020037

Dhami, M., & Ayton, P. (2001). Bailing and jailing the fast and frugal way. *Journal of Behavioral Decision Making*, *14*(2), 141–168. https://doi.org/10.1002/bdm.371

Dietvorst, B. J., Joseph, P. S., & Cade, M. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management science*, *64*(3): 1155–1170.

Dietvorst, B., Simmons, J., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dixon, H. (2020). What judges and lawyers should understand about artificial intelligence technology. *The Judges' Journal: Technology*, *59*(1), 36–38.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), Article eaao5580. https://doi.org/10.1126/sciadv.aao5580

Elhamdadi, H., Gaba, A., Kim, Y. S., & Xiong, C. (2022). How do we measure trust in visual data communication? IEEE Evaluation and Beyond-Methodological Approaches for Visualization (BELIV), Oklahoma City, OK, 17 October 2022, 85–92.

Farayola, M. M., Tal, I., Connolly, R., Saber, T., & Bendechache, M. (2023). Ethics and trustworthiness of ai for predicting the risk of recidivism: A systematic literature review. *Information*, *14*(8), 426. https://doi.org/10.3390/info14080426

Firth, A. L. (2020). Judges remain skeptical on whether artificial intelligence can make decisions more fairly than they can. *National Judicial College*. https://www.judges.org/news-and-info/judges-remain-skeptical-on-whether-artificial-intelligence-can-make-decisions-more-fairly-than-they-can/

Freeman, K. (2016). Algorithmic injustice: How the Wisconsin Supreme Court failed to protect due process rights in State v. Loomis. *North Carolina Journal of Law & Technology*, *18*(5), 75.

Fry, H. (2018). *Hello world: How to be human in the age of the machine*. Random.

Gerber, A. S., & Green, D. P. (2012). *Field experiments*. Norton.

Gerlich, M. (2023). Perceptions and acceptance of artificial intelligence: A multi-dimensional study. *Social Sciences*, *12*(9), 502. https://doi.org/10.3390/socsci12090502

Gerlich, M. (2024). Exploring motivators for trust in the dichotomy of human—AI trust dynamics. *Social Sciences*, *13*(5), 251. https://doi.org/10.3390/socsci13050251

Gino, F., & Moore, D. A. (2007). Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making*, *20*(1), 21–35. https://doi.org/10.1002/bdm.539

Gogoll, J., & Uhl, M. (2018). Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics*, *74*(June), 97–103. https://doi.org/10.1016/j.socec.2018.04.003

Goodison, S., & Brooks, C. (2023). *Local police departments, procedures, policies, and technology, 2020 – statistical tables*. U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Goodman, C. C. (2019). AI/Esq.: Impacts of artificial intelligence in lawyer-client relationships. *Oklahoma Law Review*, *72*(1), 149–184.

Greenstein, S. (2022). Preserving the rule of law in the era of Artificial Intelligence (AI). *Artificial Intelligence and Law*, *30*(3), 291–323. https://doi.org/10.1007/s10506-021-09294-4

Grossman, M. R., & Cormack, G. V. (2011). Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Law Journal of Technology*, *17*(3), 1–48. http://scholarship.richmond.edu/jolt/vol17/iss3/5

Hannah-Moffat, K. (2015). The uncertainties of risk assessment: Partiality, transparency, and just decisions. *Federal Sentencing Reporter*, *27*(4), 244–247. https://doi.org/10.1525/fsr.2015.27.4.244

Harvard Law Review. (2017). Wisconsin supreme court requires warning before use of algorithmic risk assessments in sentencing. *Harvard Law Review*, *130*(5), 1530–1537. https://www.jstor.org/stable/10.1525/fsr.2011.23.4.266

Hill, K. (2020). *Wrongfully accused by an algorithm*. The New York Times.

Ho, Y. J., Jabr, W., & Zhang, Y. (2023). AI enforcement: Examining the impact of AI on judicial fairness and public safety. *SSRN*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4533047

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. Eighth International Conference on Weblogs and Social Media (ICWSM-14), Ann Arbor, MI, 1–4 June, 2014.

Imai, K., Jiang, Z., Greiner, J., Halen, R., & Shin, S. (2023). Experimental evaluation of algorithm-assisted human decision-making: Application to pretrial public safety assessment. *Journal of the Royal Statistical Society: Series A*, *186*(2), 167–189. https://doi.org/10.1093/jrsssa/qnad010

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. *In Proceedings of the 28th European Conference on Information Systems (ECIS), An Online AIS Conference,* June 15-17, 2020, *168*. https://aisel.aisnet.org/ecis2020_rp/168

Katz, D., Michael, M., Bonmarito, J., & Blackman, J. (2017). A general approach for predicting the behavior of the Supreme Court of the United States. *PLoS One*. https://doi.org/10.2139/ssrn.2463244

Kennedy, R., Waggoner, P., & Ward, M. (2022). Trust in public policy algorithms. *The Journal of Politics*, *84*(2), 1132–1148. https://doi.org/10.1086/716283

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human decisions and machine predictions. *Quarterly Journal of Economics*, *133*(1), 237–293. https://doi.org/10.1093/qje/qjx032

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data and Society*, *5*(1), 2053951718756684. https://doi.org/10.1177/2053951718756684

Lee, M. K., & Baykal, S. (2017). Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing* (pp. 1035–1048). ACM conference. https://doi.org/10.1145/2998181.2998230

Lin, Z., Jerry, J. J., Sharad, G., & Jennifer, S. (2020). The limits of human predictions of recidivism. *Science advances*, *6*(7), eaaz0652.

Liu, H. W., Lin, C. F., & Chen, Y. J. (2019). Beyond state v Loomis: Artificial intelligence, government algorithmization and accountability. *International Journal of Law and Info Technology*, *27*(2), 122–141. https://doi.org/10.1093/ijlit/eaz001

Logg, J. M. (2016). *When do people rely on algorithms?* [Dissertation, University of California].

Logg, J. M., Minson, J., & Moore, D. A. (2018). *Algorithm appreciation: People prefer algorithmic to human judgment*. Harvard Business School NOM Unit. Working Paper No. 17–086.

Loria, S. (2018). Textblob documentation. *Release 0.15 2*, *8*, 269.

Lynch, M., & Omori, M. (2018). Crack as proxy: Aggressive federal drug prosecutions and the production of black–white racial inequality. *Law & Society Review*, *52*(3), 773–809. https://doi.org/10.1111/lasr.12348

Mac, R., Haskins, C., Sacks, B., & McDonald, L. (2021). *Your local police department might have used this facial recognition tool to surveil you. find out here*. Buzzfeed News.

Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Parli, V., Shoham, Y., Wald, R., Clark, J., & Perrault, R. (2023). Artificial intelligence index report 2023. arXiv preprint arXiv: 2310.03715. https://arxiv.org/abs/2310.03715

McDaniel, J., & Pease, K. (Eds.), (2021). *Predictive policing and artificial intelligence*. Routledge, Taylor & Francis Group.

McKnight, D. H., & Norman, L. C. (1996). *The meanings of trust*. Carlson School of Management, University of Minnesota.

Mehotra, D. (2024). *Cops used DNA to predict a suspect's face and tried to run facial recognition on it*. Wired.

Michaels, A. (2024). Elevating corporate profits over individual liberty: Comparing AI trade secret privilege in criminal proceedings with patent litigation. *Houston Law Review Online*, *33*(14). https://ssrn.com/abstract=4744905

Miron, M., Tolan, S., Gómez, E., & Castillo, C. (2021). Evaluating causes of algorithmic bias in juvenile criminal recidivism. *Artificial Intelligence and Law*, *29*(2), 111–147. https://doi.org/10.1007/s10506-020-09268-y

Mousavi Baigi, S. F., Sarbaz, M., Ghaddaripouri, K., Ghaddaripouri, M., Mousavi, A. S., & Kimiafar, K. (2023). Attitudes,

knowledge, and skills towards artificial intelligence among healthcare students: A systematic review. *Health Science Reports*, *6*(3), Article e1138. https://doi.org/10.1002/hsr2.1138

Movement Alliance Project & MediaJustice. (2024). Mapping pretrial injustice: A community-driven database. https://pretrialrisk.com/ (accessed 16 June 2024).

Okidegbe, N. (2021). Discredited data. *Cornell Law Review*, *107*, 2007–2066.

Okidegbe, N. (2022). The democratizing potential of algorithms? *Connecticut Law Review*, *53*(4), 739. https://opencommons.uconn.edu/law_review/544

O'Neill Shermer, L., Rose, K. C., & Hoffman, A. (2011). Perceptions and credibility: Understanding the nuances of eyewitness testimony. *Journal of Contemporary Criminal Justice*, *27*(2), 183–203. https://doi.org/10.1177/1043986211405886

Ozer, A. (2023). Well, you're the expert: How signals of source expertise help mitigate partisan bias. *Journal of Elections, Public Opinion, and Parties*, *33*(1), 1–21. https://doi.org/10.1080/17457289.2020.1744611

Ozer, A. L., Waggoner, P. D., & Kennedy, R. (2024). The paradox of algorithms and blame on public decision-makers. *Business and Politics*, *26*(2), 200–217. https://doi.org/10.1017/bap.2023.35

Poursabzi-Sangdeh, F., Daniel, G. G., Jake, M. H., Jennifer, W. V., & Hanna, W. (2021). Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (p. 237). New York: Association for Computing Machinery. https://doi.org/10.1145/3411764.3445315

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations*.

Rademacher, T. (2020). Artificial intelligence and law enforcement. In T. Wischmeyer & T. Rademacher (Eds.), *Regulating artificial intelligence*: Springer.

Ramakrishnan, N., Butler, P., Muthiah, S., Self, N., Khandpur, R., Saraf, S., & Wang, W. (2014). Beating the news' with EMBERS: Forecasting civil unrest using open source indicators. In *Proceedings of the ACM, SIGKDD international conference on knowledge discovery and data mining* (pp. 1799–1808).

Responsible AI Collaborative. (2024). AI incident database. https://incidentdatabase.ai/cite/74/#r1543 (accessed June 5, 2024).

Rich, B. (2018). *How AI is changing contracts*. Harvard Business Revue.

Rigano, C. (2019). *Using artificial intelligence to address criminal justice needs*. National Institute of Justice. https://www.ojp.gov/pdffiles1/nij/252038.pdf (accessed 22 May 2024).

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Rodger, H., Lensen, A., & Betkier, M. (2023). Explainable artificial intelligence for assault sentence prediction in New Zealand. *Journal of the Royal Society of New Zealand*, *53*(1), 133–147. https://doi.org/10.1080/03036758.2022.2114506

Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach*. Prentice Hall.

Ryberg, J., & Roberts, J. V. (Eds.), (2022). *Sentencing and artificial intelligence*. Oxford University Press.

Sachoulidou, A. (2023). Going beyond the "common suspects": To be presumed innocent in the era of algorithms, big data and artificial intelligence. *Artificial Intelligence and Law*, 1–54. https://doi.org/10.1007/s10506-023-09347-w

Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war*. Norton.

Segal, J., & Spaeth, H. (2002). *The Supreme Court and the attitudinal model revisited*. Cambridge University Press.

Simmler, M., Brunner, S., Canova, G., & Schedler, K. (2023). Smart criminal justice: Exploring the use of algorithms in the Swiss criminal justice system. *Artificial Intelligence and Law*, *31*(2), 213–237. https://doi.org/10.1007/s10506-022-09310-1

Stevenson, M. T., & Doleac, J. L. (2019). *Algorithmic risk assessment in the hands of humans*. IZA-Institute of Labor Economics.

Surden, H. (2021). Ethics of AI in law: Basic questions. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *In The oxford handbook of ethics of AI* (pp. 719–736). Oxford University Press.

Suroweicki, J. (2005). *The wisdom of crowds*. Anchor.

Taber, C., Cann, D., & Kucsova, S. (2009). The motivated processing of political arguments. *Political Behavior*, *31*(2), 137–155. https://doi.org/10.1007/s11109-008-9075-8

Taylor, I. (2023). Justice by algorithm: The limits of AI in criminal sentencing. *Criminal Justice Ethics*, *42*(3), 193–213. https://doi.org/10.1080/0731129x.2023.2275967

Thurman, N., Moeller, J., Helberger, N., & Trilling, D. (2019). My friends, editors, algorithms, and I: Examining audience attitudes to news selection. *Digital Journalism*, *7*(4), 447–469. https://doi.org/10.1080/21670811.2018.1493936

Tiede, L. (2007). Delegating discretion: Quasi-experiments on district court decision-making. *American Politics Research*, *35*(5), 595–620. https://doi.org/10.1177/1532673x07299449

Tiede, L. (2009). The swinging pendulum of sentencing reform: Political actors regulating district court discretion. *Brigham Young University Journal of Public Law*, *24*(1), 1–47. https://digitalcommons.law.byu.edu/jpl/vol24/iss1/2

Tyler, T. R. (2006). *Why people obey the law*. Princeton University Press.

Van Noorden, R. (2020). The ethical questions that haunt facial-recognition research. *Nature*, *587*(7834), 354–358. https://doi.org/10.1038/d41586-020-03187-3

Wagner, A. R., Borenstein, J., & Howard, A. (2018). Overtrust in the robotic age. *Communications of the ACM*, *61*(9), 22–24. https://doi.org/10.1145/3241365

Ward, J. (2019). *10 things judges should know about AI*. Judicature.

Wells, G. L. (1988). *Eyewitness identification*. Carswell.

White, M., & Eiser, J. (2010). A Social Judgement Analysis of Trust: People as Intuitive Detection Theorists 1. In S. Michael, C. E. Timothy, & G. Heinz (Eds.), *In Trust in Risk Management* (pp. 95–116). London: Routledge. https://doi.org/10.4324/9781849776592-12

Wooldredge, J., Griffin, T., & Rauschenberg, F. (2005). (Un) anticipated effects of sentencing reform on the disparate treatment

of defendants. *Law & Society Review*, *39*(4), 835–874. https://doi.org/10.1111/j.1540-5893.2005.00246.x

Xiong, C., Padilla, L., Grayson, K., & Franconeri, S. (2019). Examining the components of trust in map-based visualizations. In D. Fellner (Ed.), *Trustvis 2019 -EuroVis workshop on trustworthy visualization* (pp. 19–23). Porto, Portugal: The Eurographics Association. https://doi.org/10.2312/trvis.20191186

Yamane, N. (2020). Artificial intelligence in the legal field and the indispensable human element legal ethics demands. *Georgetown Journal of Legal Ethics*, *33*(3), 877.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*(1), 1–13. https://doi.org/10.1016/j.obhdp.2003.08.002

Završnik, A. (2020). Criminal justice, artificial intelligence systems, and human rights. In *ERA forum* (4th ed., 20, pp. 567–583). Berlin/Heidelberg: Springer Berlin Heidelberg.

Zhang, B. (2023). Public opinion toward artificial intelligence. In J. Bullock, Y. C. Chen, J. Himmelreich, V. Hudson, A. Korinek, M. Young, & B. Zhang (Eds.), *Handbook of AI governance* (pp. 553–571). Oxford University Press.

Zhang, B., & Dafoe, A. (2020). U.S. public opinion on the governance of artificial intelligence. [Paper Presentation]. Proceedings of the 2020 AAAI/ACM conference on AI, ethics, and society, New York, NY, 7–8 February, 2020.

Zhang, C., & Ma, Y. (Eds.). (2012). *Ensemble machine learning: Methods and applications*. Springer Science and Business Media.