

Applications of Risk Science to AI Fairness Evaluation: Principles, Challenges, and Best Practices

Kyra Wilson, Sabrina Kang, Saloni Dash, Aylin Caliskan

University of Washington
Seattle, WA 98195 USA

kywi@uw.edu, skang19@uw.edu, sadash@uw.edu, aylin@uw.edu

Abstract

Scholarly work which aims to describe potential societal impacts (e.g., risks) of proliferating technology (especially related to artificial intelligence or other algorithmic systems) is likely to have an impact beyond the scientific communities it was written for, given that general society itself is a primary object of study. However, it is an open question whether the current practices of AI evaluation scholarship follow the principles and best practices established by *risk science*, which aims to systematically generate knowledge related to understanding, assessing, communicating, managing, and governing risk. In this work, we examine this in depth by conducting a literature review of scholarly works purporting to evaluate the bias or fairness of technological systems used for tasks related to hiring and employment. Through analysis of 22 common fairness evaluation metrics and studies using them, we find that most characterize the severity of bias- or fairness-related consequences but do not follow best practices to characterize the uncertainty around either the occurrence of these consequences or severity estimates. Next, we conduct a case study of fairness evaluation for an AI-mediated resume screening task and demonstrate how principles of risk science can be incorporated into such an evaluation. Finally, we propose the AI Risk Report Card, which facilitates the reporting and communication of risk assessment results to stakeholders in positions to act based on the predicted risks. The outcomes of these activities suggest that further research at the convergence of risk science and AI evaluation can lead to advancements in AI assessments of societal impact by enabling shared frameworks to evaluate and discuss AI risks both within and outside of the scientific community.

Code and datasets —

<https://github.com/kyrawilson/FairRisk>

Extended version — https://arxiv.org/a/wilson_k.1.html

1 Introduction

Artificial intelligence (AI) and other advanced algorithmic systems are being created and adopted into a wider variety of tasks and settings than ever before, and thus it is critical to understand how this proliferation will positively or negatively impact people and society-at-large before undesirable

consequences occur on a wide scale. A large program of research within scientific communities has been devoted to understanding these systems’ possible impacts, especially with respect to human values such as fairness, transparency, privacy, and trust (Rismani et al. 2025), as evidenced by the growth of communities such as AIES and FAccT within the last 10 years (Wortman Vaughan et al. 2025; Varshney et al. 2025). Following widespread trends across scientific disciplines (Bornmann 2014; Roberts 2009), the impact of publications from these venues now likely expands beyond the scientific community to policy makers, organizations, and the general public who use the research to inform their own decision making. Therefore, it is more essential than ever to ensure that scholarly publications meet the needs of people outside of the scientific community; this is especially true for scholarship engaging with potential broad societal impacts of AI systems, which inherently implicates stakeholders beyond researchers and scientists.

Although the field of AI impact evaluation and analysis is relatively new, analyzing the potential for unintended impacts or consequences is a task that has been conducted across various disciplines for many years. For example, in the 1970s, nuclear power plant experts analyzed and discussed the likelihood of catastrophic events and whether these risks could be managed (and if so, how) (Aven 2011e). Similar questions were also being asked about phenomena such as global warming, smoking, and food safety, necessitating the development of a scientific discipline which could integrate approaches and ideas about risk in a domain-general way. Thus, risk science (“the practice [providing] the most updated and justified [knowledge, statements, and beliefs] on risk analysis”) emerged in the 1980s and has been growing since (Aven 2025a). In this paper, we argue that the principles of risk analysis developed through risk science are also applicable to AI impact evaluation practices, and by incorporating them, the results of scientific work can be more accessible for stakeholders beyond AI researchers (e.g., other scholars, AI users, other decision makers, etc.) to use, understand, and integrate into their own practices.

Our aims are threefold: first, we introduce readers to the field of risk science and established principles for conducting risk analysis. Second, we conduct a systematic literature review of scientific publications evaluating the fairness of

AI systems¹ used for tasks in the hiring and employment domain to determine to what extent the evaluation and reporting practices used adhere to the standards established in risk science. We chose this domain because of the significant amount of existing scholarship and the wide variety of stakeholders outside of the scientific community who are impacted by AI's use in hiring and employment (e.g., job seekers, employers, government regulators, etc.) and therefore would have an interest in understanding and potentially using the results of an AI evaluation conducted by the scientific community. These factors make this task and domain ideal for determining the extent to which best practices for risk analysis and communication are followed and highlighting areas for improvement. Finally, we conduct a risk analysis case study of an AI-mediated resume screening application to demonstrate the myriad of ways risk science principles can be applied to scholarly research in the AI fairness evaluation domain and the implications of various choices. These goals are summarized by the following findings and contributions:

1. Though risk is defined by potential consequences and the uncertainties associated with them, we find that only one out of 21 metrics commonly used to evaluate hiring tools' fairness is able to quantify probabilistic uncertainty. Furthermore, results of these evaluations were reported without any variability estimates 75% of the time, obscuring an essential component of risk characterization.
2. By re-evaluating data examining fairness and AI resume screening from Wilson and Caliskan (2024) using principles from risk science, we show how risk assessments need to incorporate both assessments of consequence severity and uncertainty to accurately and transparently characterize the risk landscape.
3. Finally, we introduce the AI Risk Report Card, which is an initial proposal for standardizing AI evaluation result reporting by incorporating key principles from risk science in order to make them more complete, understandable, and useful for a wide variety of stakeholders. Future work should empirically validate the utility of such a reporting tool and compare it to methods of communicating facts about AI systems (and the risks they pose).

2 Background

In this section, we describe the basic principles of risk science, including what risk is, how to measure and describe it, and how to communicate about it. Throughout the discussion, we offer examples of how these principles relate to current practices in AI evaluations (broadly construed, and not limited to those evaluating resume screening fairness).

2.1 The Meaning of "Risk"

The use of the word "risk" goes back hundreds of years, but as society has evolved, so has its meaning (Li, Hills, and

Hertwig 2020). In 1800, risk was strongly associated with concepts such as *danger*, *hazard*, and *fear*. While these are still associated with risk today, there are additional stronger associations with concepts such as *prevalence* and *prevention* that suggest a meaning shift away from the mere representation of an event to the ability to predict, measure, or prevent it (Li, Hills, and Hertwig 2020).

Meaning variation of risk still persists today across scientific disciplines, though at a more constrained scale than observed historical changes. For example, different definitions make different assumptions about whether risk is something that can be defined quantitatively or qualitatively; whether it exists objectively or is just derived from model concepts; if it constitutes harm alone or both harm and positive outcomes; and if it is comprised of events, uncertainties, probabilities, or some combination of these (Aven 2012). The perspective we take in this work follows from holistic definitions used in risk science—namely, that risk is a concept which exists objectively in the world, but the way that it is characterized and/or measured can be seen as inter-subjective at best (Aven, Renn, and Rosa 2011).

More specifically, risk is "uncertainty about and severity of the consequences (or outcomes) of an activity with respect to something that humans value" (Aven and Renn 2009). To illustrate the facets of this definition in the context of AI evaluations, it is a fact independent of any individual assessor that AI model behavior cannot be perfectly predicted due to the opaque and non-deterministic nature of current state-of-the-art models, and therefore uncertainty exists about the nature of consequences that may occur when using AI models. This general existence of uncertainty reflects *risk as an objective concept*. When attempts are made to characterize, assess, or attain other relevant knowledge about the unknown consequences of using AI models, then *risk as a description or measurement* is dependent (i.e., subjective or inter-subjective) on the particular circumstances of its creation. This latter kind of risk is the more relevant to creators or users of AI evaluations and, unless specified otherwise, is what we mean when discussing "risk."

Following from the definition above, the important aspects of risk to consider when attempting to measure or characterize it are 1) the consequences related to an event or activity and 2) the uncertainties associated with the event and consequences (Aven 2011e). The consequences themselves are usually described in terms of the magnitude or severity of the impact they have on something that humans value. For example, using an untested diagnostic AI model on one patient vs. many patients could have different associated risks because the potential loss of something that humans value (e.g., health, well-being, life) is much greater when more people are exposed to the model. Since the true consequence(s) or outcome(s) cannot be known in advance of their occurrence, it is also necessary to describe the uncertainties that exist surrounding potential consequences to fully characterize the risk. Figure 1 shows an illustration of this based on disparate impact in employment, which states that an unfair outcome occurs if any group is selected at less than 80% of the rate of the group with the highest selection rate (Fabris et al. 2025). Even when the average severity

¹We take an expansive view of what constitutes "AI". As noted elsewhere (Bender and Hanna 2025 - 2025), there is no consensus on definitions of AI, so we include any algorithmic artifact designed to simplify or complete tasks historically done by humans.

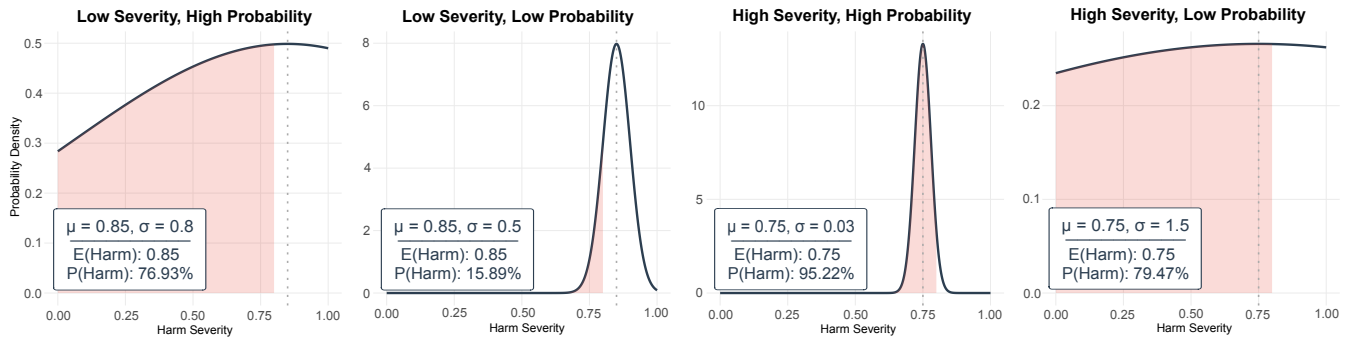


Figure 1: Illustration of how both the expected severity of a consequence and associated uncertainties play a role in risk assessment. In the two leftmost panels, the expected severity μ (0.85) is identical, but the probability that an event is harmful (any severity less than 0.8) is over 60% greater when the variability σ increases. If the expected severity falls within the range of harmful consequences (as in the two rightmost panels where severity is 0.75), increasing the variability has the opposite effect and decreases the probability of a severe outcome. Thus, considering both the severity and uncertainty or variability of possible consequence(s) is essential to understanding and effectively controlling or mitigating it.

of disparate impact stays constant, changing the uncertainty surrounding the expected harm of the outcome can vastly change the overall risk associated with a particular scenario.

The differences illustrated in Figure 1 reflect *aleatory/stochastic* uncertainties, which are variations or randomness among units that are similar to the one being studied but not observed directly (Aven 2011c). This kind of uncertainty is likely familiar to many quantitative researchers who use inferential statistics methodologies, in which probability theory is used to draw conclusions about the most likely characteristics of a population using only observations from a small sample (Thekdi and Aven 2024b). Another type of uncertainty that is necessary to characterize risk is *epistemic uncertainty*, which reflects a more general lack of knowledge about phenomena or the world and its processes (Aven et al. 2014). Crucially, epistemic uncertainty can be reduced or eliminated by gaining more knowledge about the world, while stochastic uncertainty can never be fully removed. Returning to the example in Figure 1, 0.8 was chosen as a threshold for disparate impact unfairness based on legal precedent and tradition (Fabris et al. 2025), but there is no empirical evidence (currently) that this threshold meaningfully represents unfairness better than an alternative (e.g., 0.7 or 0.9) might.

The lack of knowledge about what the “best” choices to model unfairness in the employment context reflects epistemic uncertainties that naturally exist across every domain and task. However, characterizing these uncertainties in any risk assessment is necessary to interpreting its results—for example, if epistemic uncertainty is not specified, then decision makers might make ill-informed choices based on information with only weak support (Thekdi and Aven 2024a). Alternatively, if epistemic certainty is known, it is clear whether uncertainty in the risk assessment can be decreased through further information gathering and analysis or if it is entirely stochastic and irreducible (Aven 2011d).

2.2 Assessing Consequences and Uncertainties

The first step of assessing risk for a given activity is identifying potential credible consequences or outcomes that could occur, usually with the involvement of domain experts. The initial list of possibilities may include outcomes that have no record of previously occurring, but these should still be included to ensure that consequences which are rare but high severity are not overlooked (Aven 2015; Ostrom and Wilhelmsen 2019). Once a set of possibilities is identified, the most important (with respect to causing the greatest loss or harm to things people value) can be characterized and analyzed further. In this analysis, assessors may look to historical data or testable models of systems to determine the possible ranges of consequence severities or magnitudes. In AI evaluation, this could mean reviewing reports of past harmful events archived in places like the AI Risk Repository (Slattery et al. 2026) or modeling the usage of AI models in the real world through paradigms like benchmarking, red-teaming, or auditing (Solaiman et al. 2025). Throughout all of these, descriptive statistics (e.g., identifying the range of possible consequence magnitudes and severities) are important to characterizing consequences quantitatively.

Once the severity and/or magnitude of consequences has been identified descriptively, it is necessary to identify the most likely outcomes and the level of (un)certainty about this judgment (i.e., stochastic uncertainty). As mentioned above, statistical inference can play a critical role here, as it enables assessors to use the distribution of outcome possibilities to draw conclusions about what values unobserved samples would be likely to take (Aven 2011e). For example, if a researcher was interested in whether an AI system is able to accurately diagnose a disease, they might use the values observed in a random sample of correct and incorrect AI diagnoses to determine what the theoretical expected value (i.e., center of mass in a probability distribution) of accuracy is, how much this value could vary in unobserved samples using confidence intervals, or how likely it is that the system performs better than chance or experts using null hypothesis

significance testing. All of these quantities can be derived using only frequencies (and associated frequentist probabilities) from sample observations (Aven 2011e).

However, frequentist probabilities are insufficient to describe epistemic uncertainty, and risk scientists often suggest using *subjective probabilities* instead (Aven 2011a). Under these frameworks, probability is an expression of uncertainty as perceived by an assessor using evidence such as statistical assessments, direct experience, models, or other theoretical approximations (Aven and Renn 2009). One of the most well known subjective probability paradigms is Bayesian probability, which relies on an assessor's prior knowledge of an event to characterize its probability; when the assessor receives new information, their priors are updated to accommodate this into the probability model (Aven 2020). Other methods that have been suggested include imprecise or interval probabilities, fuzzy probabilities, possibility theory, and evidence theory (Aven 2011a).

For simplicity, we will discuss epistemic uncertainty using interval probabilities, which are statements made using ranges of possible probabilities rather than a single value. For instance, rather than saying a particular AI model has a 40% chance of producing inaccurate information, an evaluator could quantify their epistemic uncertainty by saying the chance is between 25% and 50%. The interpretation of such a subjective probability statement is not the same as typical frequentist probability, but rather should be interpreted as a statement about an assessor's beliefs given the knowledge available to them. It is often helpful to compare these subjective probabilities to reference events which have commonly accepted interpretations, such as flipping coins or rolling dice (Aven and Reniers 2013). In this instance, the assessor is saying that they believe the uncertainty around inaccurate information being produced is at least the uncertainty of getting two heads on two consecutive coin flips and at most the uncertainty of getting one head in one flip.

Finally, any uncertainty estimations should also be qualified with descriptions of the knowledge used to derive them and its associated strength. The aspects of knowledge that are relevant are the same as those which would be included in any other scientific inquiry—for example, methods and materials used to conduct any empirical investigations, data analysis procedures, key assumptions, related information from outside sources, and theoretical frameworks, among many others. The choices made to generate knowledge can then be qualitatively categorized as producing strong, moderate, or weak knowledge, using the criteria proposed by Flage and Aven (2009) (reproduced in the appendix in the extended version of the paper, for example, or other similar qualitative assessment tools (see Askeland, Flage, and Aven (2017)). Importantly, the strength of knowledge is primarily a tool for decision makers or other stakeholders outside of the assessment team to contextualize and apply the results of the risk assessment for their needs (e.g., whether to gather more data, make an investment, proceed with or change a proposed plan, etc.), and it is not a reflection of the overall quality of the risk assessment (Aven 2011b). A risk assessment which uses sound methods to identify potential consequences and associated uncertainties but has limited or

unreliable data to examine would be categorized as weak by Askeland, Flage, and Aven (2017)'s criteria, but its quality could be considerably higher than one who has more reliable data available but is less transparent about the assumptions and methods used. Ultimately, the knowledge about risk and the way it is used to derive consequences and their associated uncertainties can be judged as any other scientific activities would be: through the lens of validity (i.e., how faithfully the real-world phenomena of interest are represented) and reliability (i.e., how consistent results are when the investigation is repeated) (Aven and Heide 2009).

2.3 Perceiving and Communicating Risk

The final step of risk analysis for the assessor is to communicate the findings and results to people who were not part of the assessment team. While identifying the best ways of communicating risk is still an open research area (Aven 2025b), there is some consensus that communicating probabilities shared by strength of knowledge judgments is preferable to sharing probabilities with other kinds of uncertainty judgments because they are more informative, trustworthy, and justifiable (Thekdi and Aven 2025). However, despite experts and risk analysts' best efforts to communicate risks accurately and faithfully, lay people (including those who might be ultimately responsible for using the risk assessments to make decisions) may have different understandings of the risks presented. For example, people may perceive something to be more or less risky based on how familiar they are with the particular circumstances under consideration or how much control they believe they have over the circumstances or consequences under consideration (Aven 2019). Ultimately, further research into people's risk literacy will be necessary to establish additional best practices beyond those mentioned here (Aven 2024).

The necessity of high-quality communication of AI evaluation results is also an active area of research. Inspired by model and data cards (Mitchell et al. 2019; Pushkarna, Zaldivar, and Kjartansson 2022), which aim to succinctly summarize important information about the development and intended uses of opaque technological artifacts for stakeholders outside of development teams, multiple variations of AI impact cards have been proposed for similar purposes; risk communication manifests differently in each of these. For example, (Gursoy and Kakadiaris 2022), (Stauffer et al. 2025), Bogucka et al. (2025) and state that including information about risk on their AI impact cards is important, but they provide very little information about how to describe various dimensions of risks or how risk communication differs from communicating other aspects like model capabilities. Golpayegani et al. (2024) design AI impact cards specifically based on the EU AI Act, which explicates possible risks associated with AI use in detail (Novelli et al. 2024). They include information about risk likelihood and severity, but provide little information about associated uncertainties that are essential for risk analysis and communication, as discussed in Section 2.2. Future work is needed at the intersection of risk and AI literacy in order to achieve high-quality risk communication in this domain.

3 Risk in AI Hiring Fairness Evaluations

In this section, we describe a systematic analysis of metrics used to evaluate AI tools used for hiring or employment tasks to determine the extent to which the principles of risk science are already incorporated into AI evaluation practices. We focus on evaluations of model fairness conducted primarily by and for the scientific community (i.e., those published at scholarly venues rather than as part of business or journalistic reporting). The results of this survey can elucidate which aspects of risk researchers are already measuring and describing as part of their scientific inquiries, which aspects are currently underrepresented in published scholarship and, accordingly, how this impacts people outside of the scientific community’s ability to use scientific inquiry as part of their risk assessment and decision making.

3.1 Methods

We began our review of metrics by annotating the metrics listed in Fabris et al. (2025), a multidisciplinary survey of fairness and bias in algorithmic hiring. By conducting a systematic literature review, they identify 21 unique metrics which are used for fairness evaluation across 22 different studies (on average each metric appears in 1.66 studies as some are used in multiple studies). Fabris et al. (2025) categorize each metric according to what kind of fairness it relates to (*outcome*, *impact*, *accuracy*, *representational*, *impact*, or *process*). They also annotate features of the metrics related to their interpretation and use, such as whether they allow for the inclusion of multinary social or demographic attributes or if they account for potential differences in user hyperparameter choices. We expand on these features by adding annotations for features related to risk assessment, such as whether they characterize aspects of severity or stochastic uncertainty.

For each metric, we first determine what the mathematical bounds or limits are for the metric. For example, a metric like disparate impact (which is typically formulated as the ratio between selection rates of the minority and majority groups) could have a minimum value of zero when the minority/majority group is never/always selected or a maximum value of one if the minority and majority groups have the same selection rates. Next, these bounds were used to annotate whether each metric satisfied the constraint that the probability P of any subset of events A in the set of all possible events F must be at least zero and at most one (Equation I) (Hájek 2023). For metrics that satisfied this constraint, we annotated whether they satisfied the second constraint of probability: additivity (Equation II) (Hájek 2023). This criteria states that the probabilities of all possible outcomes for a given event should sum to one exactly. Metrics which satisfied both bounds and additivity constraints were thus categorized as probability metrics (which can be used to refer to the uncertainty of a consequence), and those which did not satisfy both the constraints were categorized as severity metrics (which can characterize potential consequences).

Second, we reviewed each of the studies cited by Fabris et al. (2025) and annotated whether their experimental results using the surveyed metrics were reported as point estimates only or as point estimates with variability estimates

(e.g., confidence intervals, standard deviations, distributions, etc.). If variability estimates were included, we noted which type was used and whether these estimates were given for all results or only a subset. As no instances of formalizing epistemic uncertainties in AI evaluations were found in an initial review of the studies listed by Fabris et al. (2025), we did not include identifying or describing this type of uncertainty in our annotation procedure.

$$(I) \quad 0 \leq P(A) \leq 1 \quad \forall A \in F$$

$$(II) \quad P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

3.2 Results and Discussion

The results of our metric review and annotation are shown in Table 1. In the remainder of this section, we discuss the patterns we identified and their implications for AI fairness risk assessment and communication.

Metrics Are Dominated by Severity As shown in Table 1, the vast majority of metrics (20/21) used to assess the fairness of AI hiring tools are not valid probabilities, and therefore characterize the severity of the unfairness rather than uncertainty about whether it occurs.² The only exception to this pattern is the Sensitive Area Under the Receiver Operating Characteristic Curve (sAUC). The sAUC is a specific application of the more general Area Under the Receiver Operating Characteristic Curve (AUC-ROC), which is a function of True Positive Rates (TPR) and False Positive Rates (FPR) at various classification thresholds. As TPR and FPR can both take values from zero to one and are themselves valid probabilities, the area under any curve produced by plotting TPR and FPR against each other is also between zero and one and a valid probability (Hanley and McNeil 1982).

In the case of sAUC, the classification task is to predict a sensitive attribute (e.g., gender) using non-sensitive features that are also used in a prediction task related to hiring. The ideal outcome in this scenario is that sAUC is approximately 0.5, meaning that the classifier’s input features only afford classification at chance levels (e.g., the input features contain no information about the sensitive attributes). If sAUC is higher, it implies that there is an equivalent increase in the probability that a sensitive attribute could be correctly predicted from the input features. Therefore, the interpretation of sAUC in the AI hiring risk assessment context is that it describes the uncertainty of whether or not a person’s social or demographic identities are recoverable from other “non-sensitive” data that are used in the AI task. Other metrics would be needed to describe the severity of this consequence, such as how much or in what way this recoverability impacts the outcome of hiring or employment decisions.

The dominance of severity measures implies that important aspects of risk related to the use of AI hiring and employment tools have been overlooked in AI fairness evalu-

²Equations for each metric as well as the list of studies each was used in is shown in the appendix.

Measure	Bounds	Valid Probability
Skew@k	$(-\infty, \infty)$	No (I)
Normalized Discounted Cumulative Kullback-Leibler Divergence (NDKL)	[0,1]	No (II)
Disparate Impact (DI)	[0,1]	No (II)
Demographic Disparity (DD)	[-1,1]	No (I)
Representation in Positive Predictive Rate (RPPR)	[-1,1]	No (I)
True-Positive Rate Difference (TPRD)	[-1,1]	No (I)
False-Negative Rate Ratio (FNRR)	[0,∞)	No (I)
Extended Equality of Opportunity (ExEO)	[0,1]	No (II)
Discounted Representation Difference (DRD)	[-1,1]	No (I)
Score KL Divergence (SKLD)	[0,∞)	No (I)
Log Rank Regression (LRR)	$(-\infty, \infty)$	No (I)
Root Mean Square of TPRD (RMS)	[0,1]	No (II)
Mean Error Difference (MED)	$(-\infty, \infty)$	No (I)
Mean Absolute Error (MAE)	$(-\infty, \infty)$	No (I)
Balanced Classification Rate Difference (BCRD)	[-1,1]	No (I)
Mutual Information Difference (MID)	[-1,1]	No (I)
Salary Difference (SaD)	$(-\infty, \infty)$	No (I)
Gender Bias Score (GBS)	$(-\infty, \infty)$	No (I)
Sensitive Area Under the Receiver Operating Characteristic Curve (sAUC)	[0,1]	Yes
Ground Truth Regression (GTR)	$(-\infty, \infty)$	No (I)
Mutual Information Amplification (MIA)	[-1,1]	No (I)

Table 1: Summary of whether the AI hiring fairness metrics listed in Fabris et al. (2025) describe severity or uncertainty of harms. If a metric is not a valid probability (all but one), the constraints it violates are given in parentheses (corresponding to criteria in Equations I and II). The colors of rows correspond to the different kinds of fairness identified by Fabris et al. (2025): **Outcome**, **Accuracy**, **Impact**, **Representational**, and **Process**.

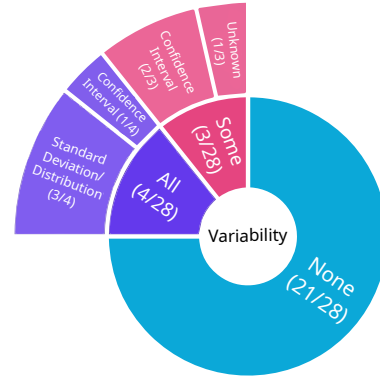


Figure 2: Among the 28 evaluations in our analysis set, only seven provided variability estimates for their metric results. Of those, four provided variability estimates for all results: one used a confidence interval, and three used a combination of standard deviations or errors and visualizations of distributions. Two papers provided confidence intervals for some but not all results, and one provided variability estimates for some results but did not specify the type.

ations up to this point. For example, while it may be important to know the expected level of disparate impact when using an AI hiring tool with no other information available, this does not provide information about how likely disparate impact (of any magnitude) would be to occur in a given situation. Both pieces are important to understanding the risk of a system in order to make decisions about it, as strategies could differ significantly based on the chance of a consequence occurring independent of its expected severity.

Results Lack Associated Variability Estimations As shown in Figure 2, across the 28 reports of results using the metrics listed in Table 1, 21 of them (75%) are given as point estimates only, without any associated measure of variability. In four examples, every time a result was given using the metrics in Table 1, it was accompanied by some form of variability estimate: one of these used confidence intervals and the other three used a combination of standard deviations or errors and visualizations of the distribution of results. In general, the visualization of full distributions is the most informative about uncertainty because they are able to communicate the overall range and shape of data as well as measures of central tendency like the mean, median, or mode. Confidence intervals and standard deviations or errors provide much less detailed information and can be difficult to interpret depending on the distributional characteristics of the observations used to produce the measures.

Communicating Epistemic Uncertainty is Challenging. Reflecting trends in other disciplines, there is no reporting standard to communicating epistemic uncertainties (Zehr 2012; Roney, Egelhofer, and Lecheler 2026), which we observed in our initial review of the 22 hiring evaluation studies. As epistemic uncertainties are a critical part of risk assessment, high-quality AI evaluations should have clear discussions of what epistemic uncertainties exist and specific

details about how they could impact the interpretation of assessed risks.

4 Case Study: Evaluating Fairness Risk of AI Hiring Tools Using Risk Science Principles

In this section, we use AI-mediated resume screening as a case study to demonstrate how the principles of risk science can be incorporated into AI fairness evaluations to improve the clarity and utility of scholarly work for stakeholders both inside and outside of the scientific community. We start by describing and discussing a previously published evaluation of fairness in AI resume screening and conclude by re-analyzing and reporting the data to better adhere to the risk science principles established previously.

4.1 Summary of Wilson and Caliskan (2024)

Resume Screening Approach and Data Before risk analysis can begin, it is necessary to have data to inform the assessments. Though many sources of data could be used depending on the goals, for this case study we chose to use data that closely matches what researchers use in typical AI fairness evaluations (e.g., outputs or behaviors elicited from AI models in experimental or observational settings) to demonstrate how methods which are already commonly used and accepted by the scientific community can be adapted to be more accessible and informative to those outside of it. Specifically, we conduct this case study using data from Wilson and Caliskan (2024), a study which investigated intersectional (gender and race) biases in resume screening conducted by large language models (LLMs).

Wilson and Caliskan (2024) used LLMs to rank hypothetical candidate resumes for a variety of open job positions. Each position was represented by a real job description that had been posted online, and resumes were also based on real resumes, though the researchers added 80 names signaling Black men, Black women, white men, and white women to the anonymous resumes to create a set of documents that represented candidates with different social identities who were equally qualified for each job. They evaluated the fairness of three LLMs (E5-mistral-7b-instruct (e5); GritLM-7B (GritLM); SFR-Embedding-Mistral (SFR)) used to rank the candidates for the job descriptions representing nine different occupations (Chief Executive, Marketing and Sales Manager, Miscellaneous Manager, Human Resource Worker, Accountant and Auditor, Miscellaneous Engineer, Secondary School Teacher, Designer, and Miscellaneous Sales and Related Worker). In total, there were approximately 9 million data points to use for fairness evaluation (comprised of 39,814 pairs of job descriptions and sets of 80 equally qualified candidates of different social identities, scored by three different LLMs).

Analysis and Results Reporting Wilson and Caliskan (2024) reported the results of their LLM evaluation at the level of model and/or occupations by averaging the fairness risk metrics computed for each combination of job description and corresponding candidate ranking. Two examples of such results are shown in Figure 3. The first result describes the severity of LLM resume screening consequences using

Risk as Reported by Wilson and Caliskan (2024)

Consequence Severity: In rankings of the top 10% of candidates, resumes with white names have an $\approx 10\%$ advantage over those with Black names when screening Accountants and Auditors with GritLM.

Consequence Uncertainty: In rankings of the top 10% of candidates, resumes with white or Black names are significantly preferred ($p < 0.05$) in 85.1% or 8.6% of tests, respectively.

Figure 3: Examples of risk characterizations given in the AI resume screening evaluation conducted by Wilson and Caliskan (2024).

a metric akin to Demographic Disparity (DD). Because the difference in selection rates is not itself a valid probability, this metric alone does not characterize any stochastic consequence uncertainty. The second result, however, does characterize uncertainty as it is based on the frequency of rankings with statistically significant disparities relative to the total number of rankings. Such a metric satisfies both the criteria in Equations I and II, and thus can characterize the stochastic uncertainty of LLM resume screening.

Each of the results presented in Figure 3 does not fully characterize unfairness risks associated with LLM resume screening. In order for the consequence severity result to be more informative of overall risk, the stochastic uncertainty of the severity should also be assessed. This would give evidence as to whether the most likely outcome is very likely (e.g., is expected to happen 90% of the time, and other possible severities have a much lower chance of occurring) or very unlikely (e.g., is expected to happen only 10% of the time, and other possible severities have only slightly smaller chances of occurring). The implications for stakeholders beyond the AI evaluators could be very different depending on how certain the prediction of the expected severity is.

In contrast, the second result in Figure 3 does characterize the stochastic uncertainty of the consequences, but it does not describe consequence severity. The statement that resumes with white names are significantly preferred indicates that a hypothesis test was conducted, and the results suggest with a high level of certainty that differences between rankings of resumes with white vs. Black names can likely be attributed to underlying differences in the way they are evaluated by the LLM and not chance (i.e., the existence of racial bias). Furthermore, these significant differences were observed in the majority of examples (85.1%). Therefore, the interpretation of this result in a predictive risk sense is that if given a set of equally qualified white and Black candidates, the chance that the white candidates would be ranked more highly than Black candidates is 85.1%. This interpretation does not give any information about how severe this ranking disparity is likely to be though—it could be relatively modest (e.g., the top 50 spots are occupied by 30 white can-

didates and 20 Black candidates) or extremely severe (e.g., the top 50 spots are occupied by all white candidates and no Black candidates). As above, the implications for stakeholders beyond the AI evaluators could also be very different depending on which level of severity is more likely.

Finally, in neither result in Figure 3 is epistemic uncertainty formalized or used to further qualify the risk estimates. This adds additional challenges to interpreting and utilizing Wilson and Caliskan (2024)’s findings, as it is unclear what different circumstances (if any) might lead to a significantly altered risk landscape. Due to the limitations identified here, the way that Wilson and Caliskan (2024) has presented their results can be significantly improved by incorporating additional characterizations of consequence severity, stochastic uncertainty, and epistemic uncertainty, in order to create a fuller assessment of the risk associated with using LLMs for resume screening.

4.2 Using Risk Science Principles to Re-Analyze Wilson and Caliskan (2024)

Describing Consequence Severity Just as Wilson and Caliskan (2024) did, our goal for this analysis is to determine the risk of using LLMs for resume screening in relation to selecting candidates which are best qualified for roles and making selections based on social identities such as race or gender. In other words, the consequence we are interested in analyzing is the selection of unequal numbers of equally qualified candidates for a given job based on their social identities, and (some of) the value(s) relevant to this consequence are justice, equality, and non-discrimination.

Using the same data as Wilson and Caliskan (2024), we first analyze the consequence of interest descriptively by calculating the Disparate Impact (DI), which is the ratio of the lowest and highest selection rates when selecting k candidates among all groups g in the set of groups S (Equation III). We chose to select the top 50 candidates as an initial point of analysis. As noted previously, the “80% rule” is a heuristic to differentiate discriminatory DI ($DI < 0.8$) from non-discriminatory ($DI \geq 0.8$), and in practice an LLM is considered fair if most scores are between 0.8 and 1; values near zero indicate high unfairness or discrimination.

$$(III) \quad DI = \frac{\min_{g \in S} N_g^k / N_g}{\max_{g \in S} N_g^k / N_g}$$

As shown in Figure 4 for the model GritLM, the average disparate impact observed across all scenarios is 0.763 (which is below the 0.8 threshold indicating that unfairness has occurred).³ However, the spread of observed values spans nearly the total range of possible DI values, with the minimum DI at 0.042 and the maximum at 1.0. This indicates that there is some amount of variability in the values that could be expected on average in new samples from the

³(Wilson and Caliskan 2024) evaluated two additional models which are omitted here for brevity. The other models showed similar patterns as illustrated in Figure 4, and these supplementary results are available in the appendix of the extended version of the paper.

same population of job descriptions and resumes. Furthermore, the distribution is left-skewed, suggesting that other descriptive measures to characterize variance (such as standard deviation) would not be appropriate in this case.

Describing Stochastic Consequence Uncertainty First, to determine what range of harm is most likely to occur on average, we conducted 5,000 iterations of bootstrap sampling to construct a 95% confidence interval showing that the average value will fall between 0.759 and 0.768 in most of the similar samples. This suggests that there is a high degree of (stochastic) certainty that the harm that is most likely is severe enough to be considered discrimination, but not the highest magnitude of discrimination possible.

While looking at the range of expected average DI values can inform stakeholders that harm is more likely to occur than not, it does not provide specific information about that likelihood. To analyze this, we code the instances where DI occurred (i.e., was less than 0.8) or not as a binary factor and sum the frequencies of each level, which is shown in Figure 4. As expected, discriminatory DI outcomes occur more often than non-discriminatory outcomes (55.41% of the time). We follow the same bootstrap sampling procedure done for raw DI values for the recoded binary values and find that the 95% confidence interval is again quite narrow, with most similar samples having harm occur in 54.11-56.70% of resume screening scenarios.

Describing Epistemic Uncertainty As demonstrated above, this evaluation has a relatively small amount of stochastic uncertainty, and so any remaining uncertainty is attributable to a lack of knowledge about the processes, applications, and scale of resume screening and discrimination in the real world. For example, as discussed in Section 2.1, the 80% rule is not empirically validated nor the only standard that can be used to determine whether or not discrimination occurs (Watkins and Chen 2024; Meier, Sacks, and Zabell 1984). In Figure 4, we show that using a different common threshold based on statistical significance⁴ can lead to very different risk characterizations.

As changing the threshold at which a consequence is considered harmful discrimination does not change the distribution of data itself, the expected severity values remain the same as when evaluating DI at the 0.8 threshold and are still most likely between 0.759 and 0.768. However, the threshold does change the interpretation of this value—at the lower threshold of 0.666, most of the observed outcomes would not be considered harmful discrimination. Therefore, the likelihood of harm occurring decreases drastically and occurs roughly 20.72% of the time (95% CI: 19.67-21.75%). This change in predicted likelihood demonstrates how characterizing epistemic uncertainty can be useful for stakeholders who may want to gather more information to reduce epistemic uncertainty before making decisions or otherwise acting on the risk assessment results. Accordingly, a possible

⁴According to the chi-square distribution, the smallest disparity that is still significant at the $\alpha = 0.05$ level with 50 candidates from two different groups is 20 vs. 30. Therefore the DI threshold based on statistical significance is $20/30 = 0.666$.

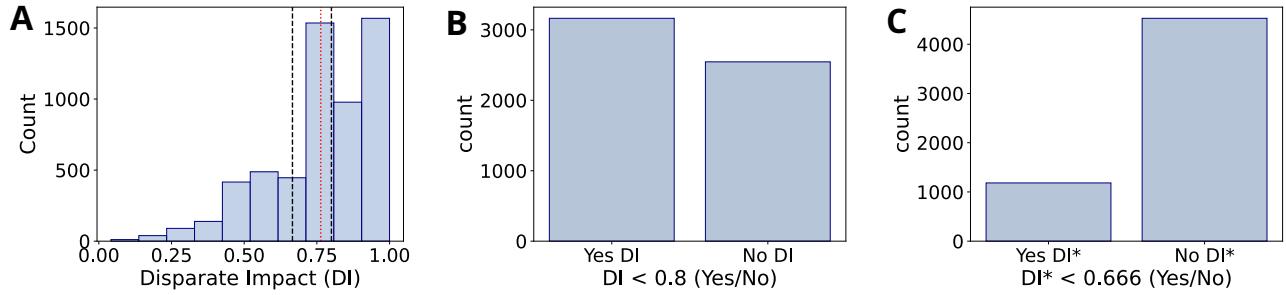


Figure 4: [A] The distribution of observed Disparate Impact (DI) scores when evaluating white men’s vs. Black men’s resumes using GritLM. The red line marks the sample average; the two dashed black lines represent different possible DI thresholds at 0.666 and 0.800. [B] The number of resume screening scenarios that are classified as discriminatory vs. non-discriminatory when the DI threshold is 0.800 (based on the 80% rule). [C] The number of resume screening scenarios which are classified as discriminatory vs. not when the DI threshold is 0.666 (based on statistical significance).

subjective probability description of AI resume screening using interval probabilities could be that harm has at least a 19.67% and at most a 56.70% chance of occurrence.

In practice, describing every possible source of epistemic uncertainty and associated changes in risk beliefs is likely impossible. For example, it would be very difficult to assess every possible DI threshold; further epistemic uncertainty also exists for whether DI is the best metric to assess fairness. Just determining the bounds of risk for all combinations of thresholds and fairness metrics is unfeasible, and yet there are still many other sources of epistemic uncertainty that could be incorporated. Thus, most current recommendations in risk science suggest characterizing epistemic uncertainties using broad qualitative judgments based on loose criteria about what constitutes “weak knowledge” vs. “strong knowledge” Askeland, Flage, and Aven (2017). For example, in this analysis, knowledge would be characterized as weak since the risk assessment involved strong simplifications and assumptions about how AI-mediated resume screening works in the real world (e.g., the LLMs used, the extent of variability in candidate resumes, the lack of human interaction or review, etc.). The full description of strength of knowledge levels and criteria is available in the appendix.

Communicating Risk: The AI Risk Report Card As noted in Aven (2025b), there is a strong need for empirical evidence on the best ways to communicate the results of risk assessments which characterize severity, stochastic uncertainty, and epistemic uncertainty. Right now, most common approaches report a combination of consequences, uncertainties, and statements about the strength of knowledge available to the evaluator used to inform their assessments (Aven 2025b). There have been a few proposals of ways to incorporate uncertainty into risk diagrams or other similar ways of visualizing risk characterizations (Goerlandt and Reniers 2016), but they are difficult to interpret and understand for non-domain experts or laypeople. Therefore, inspired by the use of model, dataset, and audit cards to communicate essential information about these technological artifacts, we propose AI Risk Report Cards as a way to com-

municate risks identified from scholarly or scientific AI evaluations to broad audiences. An example of a possible risk report card based on the re-analyzed case study of Wilson and Caliskan (2024) is shown in Figure 5.

There are six components in our AI Risk Report Card proposal. First, it is necessary to include contextual information about the **Activity** of interest (including what systems are evaluated, for what tasks, and in what domain) and the associated **Consequence** (including a description of the possible outcomes evaluated and values implicated). Next, more details about the consequence(s) should be given, including descriptions of the **Severity** and **Stochastic Uncertainty**. We suggest describing high-level results in a way that non-experts can understand (e.g., by including scale endpoints as references and minimizing the use of technical language) and summarizing these results using scales describing whether associated quantities are low, medium, high, etc. relative to other possible values. The **Epistemic Uncertainty** can be described using statements about the strength of knowledge underlying the assessment using criteria described in Flage and Aven (2009) and reproduced in the appendix. Finally, the overall **Risk Level Assessment** can be summarized by using the qualitative criteria in Aven (2017) and reproduced in the appendix as a coarse summary of the information presented elsewhere on the card. While this structure of the AI Risk Report Card was developed in order to portray the essential components of AI risk assessment and evaluation in a succinct and easily understandable manner, its utility is untested. As risk science and AI evaluation practice mature, we hope that further empirical investigation and refinement will be undertaken to validate the card and improve its efficacy.

5 Discussion

In this paper, we presented an argument that scientists and researchers investigating the potential societal impacts of AI should engage more deeply with the field of risk science, which aims to provide justified knowledge about how to conduct, use, and communicate risk analyses (Aven 2025a). As the stakeholders of AI societal impact evaluations extend far

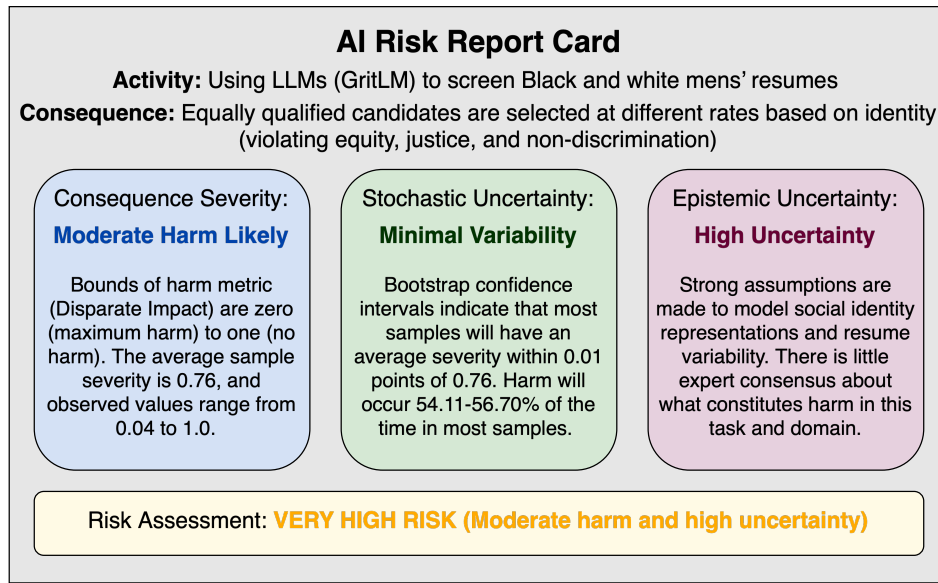


Figure 5: Example of a proposed AI Risk Report Card to facilitate communication between researchers conducting AI evaluations of societal impact and stakeholders beyond the scientific community.

beyond the scientific community, incorporating the principles of risk analysis can improve scientists' ability to make their work useful and understandable for others affected by AI proliferation. We demonstrated this in two ways: first, by conducting a literature review of fairness evaluations in AI hiring tasks and finding that most overlook critical aspects of risk such as the stochastic and epistemic uncertainty surrounding consequence characterization and measurement. Second, we demonstrated how risk principles could be incorporated into a typical AI evaluation and how this could increase the utility and interpretability of results, especially when summarized in the proposed AI Risk Report Card.

This work is in line with broader calls to improve the standards for conducting AI evaluations. For example, Stahl et al. (2023) argue that AI impact assessments are very disparate and lack shared frameworks for evaluation methods, analysis, and reporting. Risk science aims to identify generic principles of risk analysis regardless of domain, which could potentially address AI evaluation challenges related to generalizability and robustness. Another critique is of AI impact assessments is that they lack real-world grounding and validity (Blodgett et al. 2021)—while risk science cannot directly fix this, it can provide structure to characterizing and qualifying results that arise from evaluations with validity challenges through the use of epistemic uncertainty.

Future work will be needed to assess our initial proposal to incorporate risk science into AI evaluation more deeply and address its limitations, ideally opening a wide potential research agenda. For example, we situated our literature survey and case study firmly within the domain of fairness evaluation for AI hiring tasks, but it is important to determine how well our findings can generalize to other domains, tasks, and societal impacts. Because one of the primary aims of risk science is to produce general principles, we hypothe-

size it can apply to AI evaluations more broadly; future work should determine whether any evaluations are incompatible with the principles outlined here. Additionally, we looked primarily at quantitative methods of risk assessment, but additional qualitative assessment strategies can also be valuable, particularly when it comes to characterizing epistemic uncertainties (Tiusanen 2017) (and the nature of uncertainty more broadly). Finally, understanding how AI risks are perceived and how risk understanding can be facilitated through high-quality scientific risk analysis and communication is a key area for future work, which will enable greater engagement between researchers and those most directly impacted by their work throughout society.

6 Conclusion

This work represents an initial step into a larger research agenda incorporating risk science into scientific assessments of AI models' potential for positive or negative impact by motivating the necessity and benefits of such a convergence using a literature review and case study of fairness evaluation of AI used for hiring or employment tasks. We show that current practices overlook key aspects of risk such as stochastic and epistemic uncertainties, leading to incomplete and potentially misleading results. We offer suggestions for how to use risk science to improve both the practice of AI fairness evaluation and the communication of its results via the AI Risk Report Card, with the intention that understanding risk better will afford better strategies to control and mitigate risk and harm. As the societal impact of AI is something that concerns audiences far beyond the scientific community, it is critical to perform high-quality evaluations and effectively communicate their results so laypeople can incorporate the information into their own practices and decisions surrounding AI.

Acknowledgments

We are grateful to the anonymous reviewers for their constructive feedback. This work was supported in part by the Schmidt Sciences Award on AI and Advanced Computing, through the Science of Trustworthy AI program; the U.S. National Science Foundation (NSF) CAREER Award 2337877; and the University of Washington Tech Policy Lab. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of supporters.

References

- Ali, M.; Sapiezynski, P.; Bogen, M.; Korolova, A.; Mislove, A.; and Rieke, A. 2019. Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes. *Proceedings of the ACM on human-computer interaction*, 3(CSCW): 1–30.
- Arafan, A. M.; Graus, D.; Santos, F. P.; Beauxis-Aussalet, E.; et al. 2022. End-to-End Bias Mitigation in Candidate Recommender Systems with Fairness Gates. In *HR@ RecSys*.
- Askeland, T.; Flage, R.; and Aven, T. 2017. Moving beyond probabilities—Strength of knowledge characterisations applied to security. *Reliability Engineering & System Safety*, 159: 196–205.
- Aven, T. 2011a. Interpretations of alternative uncertainty representations in a reliability and risk analysis context. *Reliability Engineering & System Safety*, 96(3): 353–360.
- Aven, T. 2011b. *Misconceptions of risk*. John Wiley & Sons.
- Aven, T. 2011c. On different types of uncertainties in the context of the precautionary principle. *Risk Analysis: An International Journal*, 31(10): 1515–1525.
- Aven, T. 2011d. On some recent definitions and analysis frameworks for risk, vulnerability, and resilience. *Risk Analysis: An International Journal*, 31(4): 515–522.
- Aven, T. 2011e. *Quantitative risk assessment: the scientific platform*. Cambridge university press.
- Aven, T. 2012. The risk concept—historical and recent development trends. *Reliability Engineering & System Safety*, 99: 33–44.
- Aven, T. 2015. Implications of black swans to the foundations and practice of risk assessment and management. *Reliability Engineering & System Safety*, 134: 83–91.
- Aven, T. 2017. Improving risk characterisations in practical situations by highlighting knowledge aspects, with applications to risk matrices. *Reliability Engineering & System Safety*, 167: 42–48.
- Aven, T. 2019. *The science of risk analysis: Foundation and practice*. Routledge.
- Aven, T. 2020. Bayesian analysis: Critical issues related to its scope and boundaries in a risk context. *Reliability Engineering & System Safety*, 204: 107209.
- Aven, T. 2024. Risk literacy: Foundational issues and its connection to risk science. *Risk Analysis*, 44(5): 1011–1020.
- Aven, T. 2025a. *A History of Risk Science: Concepts and Principles of Risk Analysis*. Taylor & Francis.
- Aven, T. 2025b. An uncertainty-based risk perspective on risk perception and communication: Opportunities for new empirical-based research. *Risk Analysis*, 45(8): 2073–2078.
- Aven, T.; Baraldi, P.; Flage, R.; and Zio, E. 2014. *Uncertainty in risk assessment: the representation and treatment of uncertainties by probabilistic and non-probabilistic methods*. John Wiley & Sons.
- Aven, T.; and Heide, B. 2009. Reliability and validity of risk analysis. *Reliability Engineering & System Safety*, 94(11): 1862–1868.
- Aven, T.; and Reniers, G. 2013. How to define and interpret a probability in a risk and safety setting. *Safety science*, 51(1): 223–231.
- Aven, T.; and Renn, O. 2009. On risk defined as an event where the outcome is uncertain. *Journal of risk research*, 12(1): 1–11.
- Aven, T.; Renn, O.; and Rosa, E. A. 2011. On the ontological status of the concept of risk. *Safety Science*, 49(8-9): 1074–1079.
- Bender, E. M.; and Hanna, A. 2025 - 2025. The AI con : how to fight big tech's hype and create the future we want.
- Blodgett, S. L.; Lopez, G.; Olteanu, A.; Sim, R.; and Wallach, H. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1004–1015.
- Bogucka, E. P.; Constantinides, M.; Šćepanović, S.; and Quercia, D. 2025. Impact Assessment Card: Communicating Risks and Benefits of AI Uses. *Proceedings of the ACM on Human-Computer Interaction*, 9(7): 1–42.
- Booth, B. M.; Hickman, L.; Subburaj, S. K.; Tay, L.; Woo, S. E.; and D'Mello, S. K. 2021. Bias and fairness in multimodal machine learning: A case study of automated video interviews. In *Proceedings of the 2021 international conference on multimodal interaction*, 268–277.
- Bornmann, L. 2014. Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of informetrics*, 8(4): 895–903.
- Burke, I.; Burke, R.; and Kuljanin, G. 2021. Fair Candidate Ranking with Spatial Partitioning: Lessons from the SIOP ML competition. In *HR@ RecSys*.
- Chen, L.; Ma, R.; Hannák, A.; and Wilson, C. 2018. Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 chi conference on human factors in computing systems*, 1–14.
- De-Arteaga, M.; Romanov, A.; Wallach, H.; Chayes, J.; Borgs, C.; Chouldechova, A.; Geyik, S.; Kenthapadi, K.; and Kalai, A. T. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, 120–128.
- Deshpande, K. V.; Pan, S.; and Foulds, J. R. 2020. Mitigating demographic bias in AI-based resume filtering. In *Adjunct publication of the 28th ACM conference on user modeling, adaptation and personalization*, 268–275.

- Fabris, A.; Baranowska, N.; Dennis, M. J.; Graus, D.; Hacker, P.; Saldivar, J.; Zuiderveen Borgesius, F.; and Biega, A. J. 2025. Fairness and bias in algorithmic hiring: A multidisciplinary survey. *ACM Transactions on Intelligent Systems and Technology*, 16(1): 1–54.
- Flage, R.; and Aven, T. 2009. Expressing and communicating uncertainty in relation to quantitative risk analysis. *Reliability: Theory & Applications*, 4(2-1 (13)): 9–18.
- Geyik, S. C.; Ambler, S.; and Kenthapadi, K. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, 2221–2231.
- Goerlandt, F.; and Reniers, G. 2016. On the assessment of uncertainty in risk diagrams. *Safety Science*, 84: 67–77.
- Golpayegani, D.; Hupont, I.; Panigutti, C.; Pandit, H. J.; Schade, S.; O’Sullivan, D.; and Lewis, D. 2024. AI cards: towards an applied framework for machine-readable AI and risk documentation inspired by the EU AI Act. In *Annual Privacy Forum*, 48–72. Springer.
- Gursoy, F.; and Kakadiaris, I. A. 2022. System cards for AI-based decision-making for public policy. *arXiv preprint arXiv:2203.04754*.
- Hanley, J. A.; and McNeil, B. J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1): 29–36.
- Hemamou, L.; and Coleman, W. 2022. Delivering fairness in human resources AI: Mutual information to the rescue. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 867–882.
- Hemamou, L.; Guillon, A.; Martin, J.-C.; and Clavel, C. 2021. Don’t judge me by my face: An indirect adversarial approach to remove sensitive information from multimodal neural representation in asynchronous job video interviews. *arXiv preprint arXiv:2110.09424*.
- Hu, S.; Al-Ani, J. A.; Hughes, K. D.; Denier, N.; Konnikov, A.; Ding, L.; Xie, J.; Hu, Y.; Tarafdar, M.; Jiang, B.; et al. 2022. Balancing gender bias in job advertisements with text-level bias mitigation. *Frontiers in big Data*, 5: 805713.
- Hájek, A. 2023. Interpretations of Probability. In Zalta, E. N.; and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition.
- Imana, B.; Korolova, A.; and Heidemann, J. 2021. Auditing for discrimination in algorithms delivering job ads. In *Proceedings of the web conference 2021*, 3767–3778.
- Köchling, A.; Riazzy, S.; Wehner, M. C.; and Simbeck, K. 2021. Highly Accurate, But Still Discriminatory: A. Köchling et al. *Business & Information Systems Engineering*, 63(1): 39–54.
- Li, Y.; Hills, T.; and Hertwig, R. 2020. A brief history of risk. *Cognition*, 203: 104344.
- Meier, P.; Sacks, J.; and Zabell, S. L. 1984. What happened in Hazelwood: Statistics, employment discrimination, and the 80% rule. *American Bar Foundation Research Journal*, 9(1): 139–186.
- Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.
- Nandy, P.; Diccio, C.; Venugopalan, D.; Logan, H.; Basu, K.; and El Karoui, N. 2022. Achieving Fairness via Post-Processing in Web-Scale Recommender Systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 715–725.
- Novelli, C.; Casolari, F.; Rotolo, A.; Taddeo, M.; and Floridi, L. 2024. Taking AI risks seriously: A new assessment model for the AI Act. *Ai & Society*, 39(5): 2493–2497.
- Ostrom, L. T.; and Wilhelmsen, C. A. 2019. *Risk assessment: tools, techniques, and their applications*. John Wiley & Sons.
- Parasurama, P.; and Sedoc, J. 2021. Degendering resumes for fair algorithmic resume screening. *arXiv preprint arXiv:2112.08910*.
- Parasurama, P.; Sedoc, J.; and Ghose, A. 2022. Gendered information in resumes and hiring bias: A predictive modeling approach. *Available at SSRN 4074976*.
- Pena, A.; Serna, I.; Morales, A.; and Fierrez, J. 2020. Bias in multimodal AI: Testbed for fair automatic recruitment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 28–29.
- Pushkarna, M.; Zaldivar, A.; and Kjartansson, O. 2022. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 1776–1826.
- Rismani, S.; Shelby, R.; Davis, L.; Rostamzadeh, N.; and Moon, A. 2025. Measuring What Matters: Connecting AI Ethics Evaluations to System Attributes, Hazards, and Harms. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8(3), 2199–2213.
- Roberts, M. R. 2009. Realizing societal benefit from academic research: Analysis of the National Science Foundation’s broader impacts criterion. *Social Epistemology*, 23(3-4): 199–219.
- Roney, C.; Egelhofer, J. L.; and Lecheler, S. 2026. The “bread and butter” of science: understanding how scientists communicate uncertainties. *Humanities and Social Sciences Communications*.
- Rus, C.; Luppens, J.; Oosterhuis, H.; and Schoenmacker, G. H. 2022. Closing the gender wage gap: Adversarial fairness in job recommendation. *arXiv preprint arXiv:2209.09592*.
- Singhania, A.; Unnam, A.; and Aggarwal, V. 2020. Grading video interviews with fairness considerations. *arXiv preprint arXiv:2007.05461*.
- Slattery, P.; Saeri, A. K.; Grundy, E. A.; Graham, J.; Noetel, M.; Uuk, R.; Dao, J.; Pour, S.; Casper, S.; and Thompson, N. 2026. The AI risk repository: A meta-review, database, and taxonomy of risks from artificial intelligence. *Patterns*.

- Solaiman, I.; Talat, Z.; Agnew, W.; Ahmad, L.; Baker, D. K.; Blodgett, S. L.; Chen, C.; Daumé, I.; Hal, Dodge, J.; Duan, I.; Evans, E.; Friedrich, F.; Ghosh, A.; Gohar, U.; Hooker, S.; Jernite, Y.; Kalluri, P. R.; Leidinger, A.; Lusoli, A.; Lin, M.; Lin, X.; Luccioni, S.; Mickel, J.; Mitchell, M.; Newman, J.; Ovalle, A.; Png, M.-T.; Singh, S.; Strait, A.; Struppek, L.; Subramonian, A.; and Vassilev, A. 2025. Evaluating the Social Impact of Generative AI Systems. In Hacker, P.; Engel, A.; Hammer, S.; and Mittelstadt, B., eds., *The Oxford Handbook of the Foundations and Regulation of Generative AI*. Oxford University Press. ISBN 9780198940272.
- Stahl, B. C.; Antoniou, J.; Bhalla, N.; Brooks, L.; Jansen, P.; Lindqvist, B.; Kirichenko, A.; Marchal, S.; Rodrigues, R.; Santiago, N.; et al. 2023. A systematic review of artificial intelligence impact assessments. *Artificial Intelligence Review*, 56(11): 12799–12831.
- Stauffer, L.; Yang, M.; Reuel, A.; and Casper, S. 2025. Audit cards: Contextualizing ai evaluations. *arXiv preprint arXiv:2504.13839*.
- Thekdi, S.; and Aven, T. 2024a. A classification system for characterizing the integrity and quality of evidence in risk studies. *Risk Analysis*, 44(1): 264–280.
- Thekdi, S.; and Aven, T. 2024b. Understanding explainability and interpretability for risk science applications. *Safety Science*, 176: 106566.
- Thekdi, S.; and Aven, T. 2025. Evaluating risk analyst views on uncertainty and knowledge aspects for risk characterization approaches. *Journal of Risk Research*, 28(8): 912–928.
- Tiusanen, R. 2017. Qualitative risk analysis. *Handbook of safety principles*, 463–492.
- Varshney, K.; Lechterman, T.; Burton, E.; Mattei, N.; Páez, A.; Palmiotto, F.; Farrell, T.; and Zhang, W. 2025. The Program Guide for the Eighth AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society.
- Watkins, E. A.; and Chen, J. 2024. The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 764–775.
- Wilson, C.; Ghosh, A.; Jiang, S.; Mislove, A.; Baker, L.; Szary, J.; Trindel, K.; and Polli, F. 2021. Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 666–677.
- Wilson, K.; and Caliskan, A. 2024. Gender, race, and intersectional bias in resume screening via language model retrieval. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7(1), 1578–1590.
- Wortman Vaughan, J.; Mohamed, S.; Fazelpour, S.; and Gillis, T. B. 2025. Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency.
- Yan, S.; Huang, D.; and Soleymani, M. 2020. Mitigating biases in multimodal personality assessment. In *Proceedings of the 2020 international conference on multimodal interaction*, 361–369.
- Zehr, S. C. 2012. Scientists’ representations of uncertainty. In *Communicating uncertainty*, 3–21. Routledge.
- Zhang, S.; Li, P.; and Cai, Z. 2022. Are male candidates better than females? Debiasing BERT resume retrieval system. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 616–621. IEEE.

A Qualitative Criteria for Assessing Strength of Knowledge (Reproduced from Flage and Aven (2009))

Uncertainty	Strength of Knowledge	Criteria	Conditions
Low	Strong	All conditions are fulfilled	The phenomena involved are well understood; the models used are known to give predictions with the required accuracy Much reliable data are available The assumptions made are seen as very reasonable There is broad agreement among experts
Medium	Moderate	Conditions between low and high are fulfilled to varying degrees	Example: The phenomena involved are well understood, but the models used are considered simple Example: Some reliable data are available
High	Strong	At least one condition is fulfilled	The phenomena involved are not well understood; models are non-existent or known to give poor predictions Data are not available, or are unreliable The assumptions made represent strong simplifications There is lack of agreement/consensus among experts

B Qualitative Criteria for Relative Risk Ranking (Reproduced from Aven (2017))

Rank	Criteria
Very high risk	Potential for extreme consequences, relatively large associated probability of such consequences and/or significant uncertainty (relatively weak background knowledge)
High risk	The potential for extreme consequences, relatively small associated probability of such consequences and moderate or weak background knowledge
Moderate risk	Between low and high risk. For example, the potential for moderate consequences, and weak background knowledge.
Low risk	No potential for serious consequences

C Fairness Evaluation Equations and Associated Studies (Reproduced from Fabris et al. (2025))

Measure	Equation	Studies
Skew@k	$\text{skew@k}_g = \log \left(\frac{N_g^k / k}{D_g} \right)$	(Geyik, Ambler, and Kenthapadi 2019)
Normalized Discounted Cumulative Kullback-Leibler Divergence (NDKL)	$\text{NDKL} = \frac{1}{Z} \sum_{k=1}^K \frac{1}{\log_2(k+1)} d_{\text{KL}}(D, N^k)$	(Geyik, Ambler, and Kenthapadi 2019) (Arafan et al. 2022)
Disparate Impact (DI)	$\text{DI} = \frac{\min_{g \in \mathcal{S}} N_g^k / N_g}{\max_{g \in \mathcal{S}} N_g^k / N_g}$ $= \frac{\min_{g \in \mathcal{S}} \Pr(\hat{y}=1 s=g)}{\max_{g \in \mathcal{S}} \Pr(\hat{y}=1 s=g)}$	(Booth et al. 2021) (Burke, Burke, and Kuljanin 2021) (Köchling et al. 2021) (Wilson et al. 2021) (Deshpande, Pan, and Foulds 2020)
Demographic Disparity (DD)	$\text{DD} = N_g^k / N_g - N_{g^c}^k / N_{g^c}$ $= \Pr(\hat{y} = 1 s = g) - \Pr(\hat{y} = 1 s = g^c)$	(Rus et al. 2022)
Representation in Positive Predictive Rate (RPPR)	$\text{RPP}_g = \Pr(s = g \hat{y} = 1)$ $\text{RPPD} = \text{RPP}_g - \text{RPP}_{g^c} = 2 \text{RPP}_g - 1$	(Ali et al. 2019) (Imana, Korolova, and Heidemann 2021) (Pena et al. 2020)
True-Positive Rate Difference (TPRD)	$\text{TPR}_g = \Pr(\hat{y} = 1 y = 1, s = g)$ $\text{TPRD} = \text{TPR}_g - \text{TPR}_{g^c}$	(De-Arteaga et al. 2019) (Hemamou and Coleman 2022)
False-Negative Rate Ratio (FNRR)	$\text{FNR}_g = \Pr(\hat{y} = 0 y = 1, s = g)$ $\text{FNRR} = \frac{\text{FNR}_g}{\text{FNR}_{g^c}}$	(Köchling et al. 2021)
Extended Equality of Opportunity (ExEO)	$\text{xEO} = \max_x [\Pr(f_{\text{soft}}(x) \leq t y = 1, s = g) - \Pr(f_{\text{soft}}(x) \leq t y = 1, s = g^c)]$	(Nandy et al. 2022)
Discounted Representation Difference (DRD)	$\text{DRD} = \sum_{k=1}^K \frac{1}{\log_2(k+1)} [\mathbb{1}(s_{\tau(k)} = g) - \mathbb{1}(s_{\tau(k)} = g^c)]$	(Zhang, Li, and Cai 2022)
Score KL Divergence (SKLD)	$\text{SKL} = \text{KL}(D_g^f, D_{g^c}^f)$	(Pena et al. 2020)
Log Rank Regression (LRR)	$\log(\tau^{-1}(i)) = \beta_x x_i + \beta_s s_i + \mu + \epsilon$ $\text{LRR} = \hat{\beta}_s$	(Chen et al. 2018)
Root Mean Square of TPRD (RMS)	$\text{RMS} = \sqrt{\frac{1}{ \mathcal{S} } \sum_{g \in \mathcal{S}} \text{TPRD}_g^2}$	(Hemamou and Coleman 2022)
Mean Error Difference (MED)	$\text{MED} = \max_{g \in \mathcal{S}} \frac{1}{N_g} \sum_{i \in g} (y_i - f_{\text{soft}}(x_i)) - \min_{g \in \mathcal{S}} \frac{1}{N_g} \sum_{i \in g} (y_i - f_{\text{soft}}(x_i))$	(Singhania, Unnam, and Aggarwal 2020)
Mean Absolute Error (MAE)	$\text{MAE} = \frac{1}{N_g} \sum_{i \in g} y_i - f_{\text{soft}}(x_i) - \frac{1}{N_{g^c}} \sum_{i \in g^c} y_i - f_{\text{soft}}(x_i) $	(Yan, Huang, and Soleymani 2020) (Singhania, Unnam, and Aggarwal 2020)
Balanced Classification Rate Difference (BCRD)	$\text{BCR}_g = \frac{\text{TPR}_g + \text{TNR}_g}{2}$ $\text{BCRD} = \text{BCR}_g - \text{BCR}_{g^c}$	(Köchling et al. 2021)
Mutual Information Difference (MID)	$\text{MI}_g = \text{MI}_{\{i \in g\}}(\hat{y}_i, y_i)$ $\text{MID} = \text{MI}_g - \text{MI}_{g^c}$	(Köchling et al. 2021)
Salary Difference (SaD)	$f(x_i) : \text{closest job match}; W(f(x_i)) : \text{avg wage}$ $\text{SD} = \frac{1}{N_g} \sum_{i \in g} W(f(x_i)) - \frac{1}{N_{g^c}} \sum_{i \in g^c} W(f(x_i))$	(Rus et al. 2022)
Gender Bias Score (GBS)	$\text{GBS} = \text{sign}(x_m - x_f) \cdot \max \left\{ \frac{x_m - x_f}{x_m}, \frac{x_f - x_m}{x_f} \right\}$	(Hu et al. 2022)
Sensitive Area Under the Receiver Operating Characteristic Curve (sAUC)	$\text{sAUC} = \text{AUC}(h_{\text{soft}}(x))$	(Parasurama, Sedoc, and Ghose 2022) (Parasurama and Sedoc 2021) (Booth et al. 2021) (Hemamou et al. 2021) (Rus et al. 2022) (Hemamou and Coleman 2022)
Ground Truth Regression (GTR)	$\log \left(\frac{y_i}{1-y_i} \right) = \beta_x x_i + \beta_s s_i + \mu + \epsilon$ $\text{GTR} = \beta_s$	(Parasurama, Sedoc, and Ghose 2022)
Mutual Information Amplification (MIA)	$\text{MIA} = \text{MI}(\hat{y}, s) - \text{MI}(y, s)$	(Yan, Huang, and Soleymani 2020)

Table 2: Summary of the AI hiring fairness metrics listed in Fabris et al. (2025) and the studies which used them. The colors of rows correspond to the different kinds of fairness identified by Fabris et al. (2025): Outcome, Accuracy, Impact, Representational, and Process.

D Supplemental Results

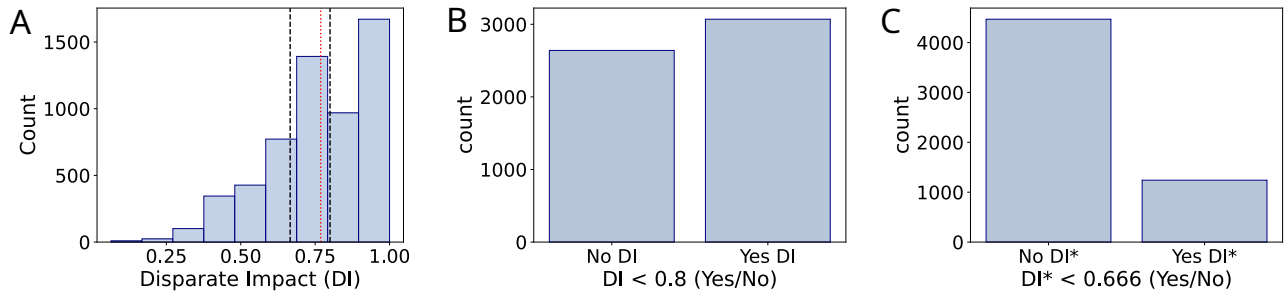


Figure 6: [A] The distribution of observed Disparate Impact (DI) scores when evaluating white men's vs. Black men's resumes using e5. The red line marks the sample average; the two dashed black lines represent different possible DI thresholds at 0.666 and 0.800. [B] The number of resume screening scenarios that are classified as discriminatory vs. non-discriminatory when the DI threshold is 0.800 (based on the 80% rule). [C] The number of resume screening scenarios which are classified as discriminatory vs. not when the DI threshold is 0.666 (based on statistical significance).

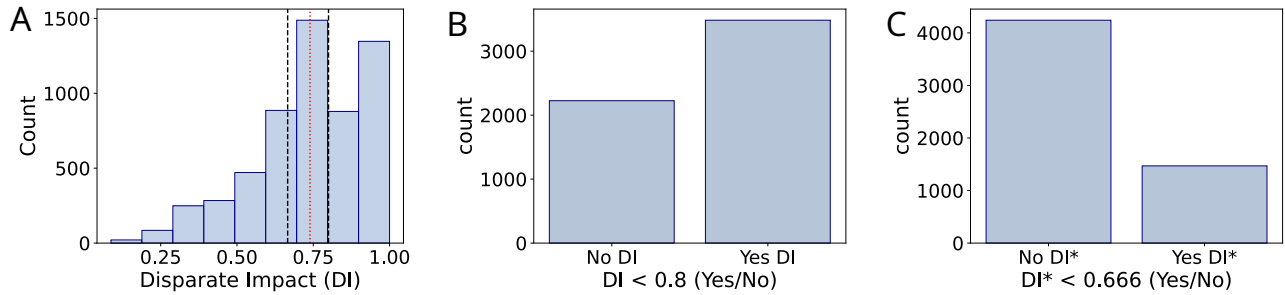


Figure 7: [A] The distribution of observed Disparate Impact (DI) scores when evaluating white men's vs. Black men's resumes using SFR. The red line marks the sample average; the two dashed black lines represent different possible DI thresholds at 0.666 and 0.800. [B] The number of resume screening scenarios that are classified as discriminatory vs. non-discriminatory when the DI threshold is 0.800 (based on the 80% rule). [C] The number of resume screening scenarios which are classified as discriminatory vs. not when the DI threshold is 0.666 (based on statistical significance).