Appendices and Supplemental Materials

Appendix A (Section 3.1.3): PRISMA Scoping Review Diagram

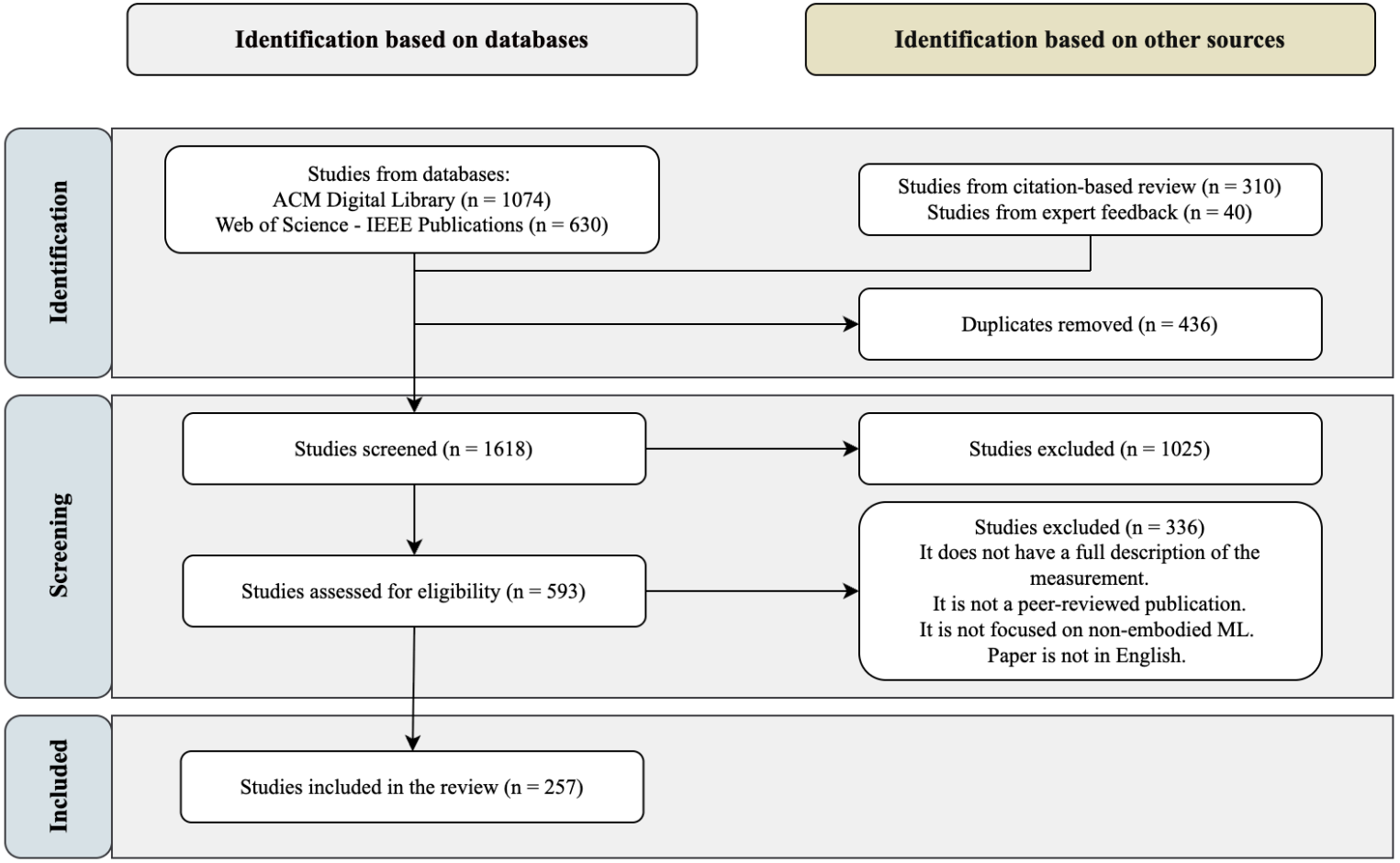

Figure A1: An ordered summary of the scoping review process followed to gather the articles included in the review.

Appendix B (Section 3.3): Dataset Features

Table B1: Description of all seventeen dataset features, including the generalized category they are categorized by, the variable name, their data type, and a textual description.

User Guidance Categories	Variable Name	Variable Type	Description
Target Output: The resulting measures collected in this dataset.	Measure	Textual	The measure name.
	Measurement Process	Textual	The measure description includes the evaluation process followed.
Entry Points: The primary features for filtering potential measures for an algorithmic system.	Principle	Categorical	The principle(s) are framed as the theoretical construct being measured in the paper e.g., fairness, privacy, or solidarity. Eleven principles are considered.
	ML System Component	Categorical	The component of an AI system that is being assessed. Five categories of Input data, model, output, interaction with the user, and full system are considered.
Connections to Harm: The harms, hazards, and attributes related to the measure and measurement process.	Primary Harm	Categorical	The primary harm that the measure is related to. It can be one of five options: representational, allocative, quality of service, interpersonal harms, and social system harms.
	Secondary Harm	Categorical	The secondary harm (if present) that the measure is related to. It can be one of five options: representational, allocative, quality of service, interpersonal harms, and social system harms.
	Hazard	Textual	Description of the potential actions or activities that may lead to harm.
	Attribute	Textual	Qualification or quantification of the measurement process taking place.
Measurement Properties: The standard(s) used in each measure's evaluation.	Criterion Name	Textual	The name of a standard by which something may be judged or decided in the measurement process.
	Criterion Description	Textual	A detailed description of the identified criterion.
	Type of Assessment	Categorical	Each measure can be described by at minimum one of five assessment types, including statistical, mathematical, behavioral, self-reported, or other.

Algorithmic System Characteristics: Additional features that a user can consider when narrowing down measures to use.	Application Area	Categorical	The general context in which this ML system is used e.g., healthcare, education, or transportation.
	Purpose of ML System	Textual	The goal/objective of the ML system.
	Type of Data	Textual	The data format and type used in training and/or evaluating the ML system.
	Algorithm Type	Textual	The type of ML algorithm.
Publication Metadata: Details further documentation into each source extracted to collect each feature and measure.	Title	Textual	The title of the paper that contains the corresponding measure. It is hyperlinked for direct article access.
	Publication Year	Categorical	The publication year of the paper.
	DOI Link	Textual	DOI link to the article for further information.

Appendix C (Section 3.3): Access to GitHub Repository, Dataset, and Deployed Web Visualization

The GitHub repository for this paper can be found at https://github.com/RAISE-Lab/Measuring-What-Matters-Connecting-AI-Ethics-Evaluations-to-System-Attributes-Hazards-and-Harms. The deployed web visualization can be found at https://rai-measures.onrender.com/. The GitHub repository contains a README file; the completed dataset with all seventeen columns (listed in a Microsoft Excel (.xlsx) format); an accessible link to the web-deployed interactive visualization; and a Python script and code package requirements text file to run the web-deployed visualization.

Appendix D (Section 4.2): Illustrative Mapping of Measures to Attributes, Hazards, and Harms

Representational Harms

Table D1: Summary of the categories, attributes, and hazards extracted within representational harms, with examples explained by various measures.

Category	Attribute	Hazard	Measure Examples
Whose identities and attributes are included or omitted in an AI system, and in what way?	**The presence of protected groups in the AI system outputs.** Examples: **A1.** Composition of the outcome based on sociodemographics **A2.** Bias at sentence level (fill-in-the-blank)	**Underrepresentation or omission of specific groups in system outputs.** Examples: **H1.** A disproportionate percentage of recommended content comes from a single sociodemographic group **H2.** The model generates text that reinforces stereotypes when completing sentences	**M1. Percentage of Content by Group: C**alculates the average percentage of recommended content attributed to a particular sociodemographic group (e.g., female artists) to assess representation in recommendation rankings. **M2. Intrasentence Context Association Test:** A fill-in-the-blank task where the model ranks the likelihood of a stereotypical, anti-stereotypical, and meaningless option completing a sentence. Measures how often the model prefers stereotypical over anti-stereotypical content.
	Representation within input datasets used to train AI systems. Examples: **A1.** Visual salience of individuals based on demographic traits (e.g., gender, skin tone) **A2.** Distribution of favorable labels in the transformed input data across sensitive groups	**Imbalanced or skewed representation in training data.** Examples: **H1.** People from marginalized groups (e.g., women, darker-skinned individuals) appear smaller or further from the center in images, signaling lower importance or focus **H2.** The preprocessing step results in unequal distributions of favorable examples across protected and unprotected groups	**M1. Person Prominence:** Measures the proportion of an image that a person occupies and their distance from the image center. **M2. Statistical Parity Difference:** Assesses the fairness of a preprocessing stage by comparing how often data instances from unprivileged and privileged groups are transformed from unfavorable to favorable outcomes. It quantifies whether the input data transformation introduces or mitigates group disparities.
	Internal model representations. Examples: **A1.** Disparities in learned visual representations of social memberships of people **A2.** Stereotypical associations	**Reinforcement of stereotypes through learned embeddings or features.** Examples: **H1.** There are significant differences in how people are represented in the embedding space based on attributes like age, gender, skin tone, and race	**M1. Cosine Similarity of Images**: Measures the similarity between image embeddings in a visual feature extractor using cosine similarity; used to assess disparity in learned representations across sensitive attributes such as gender, age, race, and skintone. **M2. Word Embedding Association Test (WEAT)**: Quantifies stereotypical associations between target

		H2. There are strong stereotypical associations in the embedding space	concepts (e.g., gender categories) and attributes (e.g., career vs. family) in word embeddings using cosine similarity.
Who receives what outcomes from the AI system, and are they distributed equitably across groups?	**Distribution of outcomes across different groups.** Examples: **A1.** Distribution of outcomes given sensitive attributes **A2.** Distribution of false positive errors across demographic groups	**Unequal allocation of beneficial or harmful outcomes.** Examples: **H1.** The outcomes are not distributed equally across protected and unprotected groups **H2.** The system incorrectly flags more negative cases as positive for one group than another	**M1. Total Average Equality**: The average of five fairness metrics—accuracy equality, statistical parity, conditional procedure accuracy, conditional use accuracy equality, and treatment equality—capturing overall disparity in outcomes across protected groups. **M2. False Positive Rate Difference:** Measures disparity in the false positive rate—the proportion of negative examples incorrectly predicted as positive—across demographic groups. Under fairness criteria like Equalized Odds, a fair system should yield similar false positive rates for all groups. The FPR difference is computed as: FPR (privileged group) − FPR (unprivileged group).
To what extent do users feel that the system's outputs reflect their perspectives or lived experiences?	**Personal relevance and cultural resonance of system outputs.** Examples: **A1.** The adaptability of the model's responses to reflect group-specific perspectives **A2.** The extent to which users are collectively familiarized with shared content categories through ranked recommendations	**Outputs that are irrelevant, unfamiliar, or misaligned with users' identities and experiences.** Examples: **H1.** The model is not able to adjust its responses based on context and instruction, resulting in poor alignment with the intended demographic viewpoint **H2.** The system's recommendations do not expose users to a sufficient and shared set of relevant content, resulting in low collective familiarity and the erosion of shared cultural understanding	**M1. Steerability:** Measures the average alignment between the model's responses and those of a specific demographic group, when the model is prompted with group-specific context (e.g., "Answer as a conservative person"). A higher steerability score indicates that the model can adapt its outputs to reflect that group's perspective. **M2. Commonality:** Measures the probability that all users simultaneously gain familiarity with a set of editorially selected content categories after engaging with a ranked list of recommendations. Familiarity is estimated using recall—i.e., the fraction of relevant content a user has encountered in a session. Rather than computing average familiarity across users (which can be skewed by outliers), commonality models it as a joint distribution over users, capturing whether a shared cultural experience is achieved across the population.

Allocative Harms

Table D2: Summary of the categories, attributes, and hazards extracted within allocative harms, with examples explained by various measures.

Category	Attribute	Hazard	Measure Examples
How are decisions, opportunities, or resources distributed across different groups?	**Distribution of outcomes.** Examples: **A1.** Distribution of outcomes given sensitive attributes **A2.** Disparity between system exposure and target exposure for individual users and items	**Certain groups receive fewer beneficial outcomes.** Examples: **H1.** The model produces significantly fewer positive outcomes for the disadvantaged group **H2.** Certain users never see relevant items while others are disproportionately shown particular content—contributing to skewed access to opportunities or information.	**M1. Disparate Impact:** Quantified as the ratio of positive predictions for the disadvantaged group to those for the advantaged group. **M2. Individual-User-To-Individual-Item-Fairness (II-F):** Evaluates how equitably a recommender system distributes exposure between individual users and individual items. For each user–item pair, the system compares the actual exposure (visibility) the item received to a target exposure grounded in fairness principles (like equal expected exposure).
	Distribution of errors. Examples: **A1.** Ratio of error types across groups	**Error rates disproportionately affect specific groups.** Examples: **H1.** Ratio of false positives and false negatives - The model's false negatives and false positives are unevenly distributed across social groups.	**M1. Treatment Equality (Ratio of FN and FP):** Treatment equality assesses whether the ratio of false negatives to false positives is the same across protected groups. It highlights whether one group is more likely to face more severe types of misclassification (e.g., wrongly denied parole) than another, signaling unequal error burdens.
How well does the system perform across different groups?	**Performance of the AI system.** Examples: **A1.** Difference in performance	**Model accuracy or reliability is significantly lower for certain groups, leading to inconsistent or incorrect outcomes.** Examples: **H1.** The system produces unequal prediction accuracy across groups, leading to unfair or biased treatment.	**M1. Difference in Acceptance and Rejection Rates:** Measures disparity in precision (acceptance rate) and rejection accuracy (rejection rate) between advantaged and disadvantaged groups. Acceptance rate is calculated as the proportion of correct positive predictions; rejection rate as correct negative predictions.

Do users perceive the system's decisions as fair and understandable?	**Perceived understanding of decision rationale.** Examples: **A1.** User's perceived understanding of the decision-making logic	**Users do not have a sufficient understanding of the decision rationale.** Examples: **H1.** Users lack a clear grasp of how and why the AI system made a decision	**M1. Understanding:** Participants rated their agreement with the statement: "I understand the process by which the decision was made," following exposure to different explanation styles for algorithmic decisions.
	Perception of fairness for the AI system. Examples: **A1.** Perceived fairness of the training process **A2.** Procedural justice: perceived appropriateness of decision factors	**Users perceive the system's decisions as biased or unjust.** Examples: **H1.** Users believe the training data was biased or opaque. **H2.** People may believe the system used irrelevant or unfair criteria, leading to perceived injustice.	**M1. Perceived Fairness of the Training Process:** Participants rate how fair they believe the system's training process was, based on a 1–7 Likert scale after viewing data-centric explanation **M2. Appropriateness of Factors:** Participants rated whether the factors considered in the AI decision were appropriate, using a 5-point Likert scale after being presented with a decision scenario and explanation.
To what extent do input features and modeling strategies influence disparities in outcomes?	**Distribution of dataset features and labels.** Examples: **A1.** Representation in the dataset **A2.** Distribution in labels	**Training data underrepresents or misrepresents certain groups.** Examples: **H1.** A disadvantaged group is poorly represented in the dataset, potentially leading to biased model performance and inequitable outcomes **H2.** The labels are disparately distributed across protected groups, potentially reinforcing existing inequalities	**M1. Class Imbalance:** Measures the representation gap in the dataset by computing the difference between advantaged and disadvantaged group counts, normalized by total instances. **M2. Conditional Disparity in Labels (CDDL):** Refines demographic disparity by stratifying the data and weighting group disparities across strata.
	Model characteristics accounting for fairness. Examples: **A1.** Losses of the model — specifically, the change in predictive loss (e.g., MSE) under fairness constraints.	**Model design choices (e.g., objective functions, regularization) fail to mitigate or exacerbate group disparities.** Examples: **H1.** The model incurs a high predictive loss when fairness constraints are enforced (or vice versa).	**M1. Price of Fairness:** A normalized metric that quantifies the cost of improving fairness by measuring the relative increase in model loss (e.g., mean squared error) required to achieve a specified reduction in fairness penalty. It captures the trade-off between fairness and accuracy by comparing the constrained model's loss to the unconstrained model's loss.

Quality of Service Harms

Table D3: Summary of the categories, attributes, and hazards extracted within quality of service harms, with examples explained by various measures.

Category	Attribute	Hazard	Measure Examples
Does the AI system perform its tasks with equal accuracy and reliability across all user groups?	**Performance of the AI system.** Examples: **A1.** Performance quality of the inference **A2.** Performance across different attributes **A3.** Distribution of exposure for different items across different groups	**Disparate performance of the AI system across groups.** Examples: **H1.** The inference has low precision for one group compared to the other, leading to disparities in content exposure **H2.** There is a significant performance disparity across protected groups, indicating the model performs better for some groups than others **H3.** There is a disparity in the exposure of recommended items for various groups, leading to allocative unfairness and reinforcing visibility imbalances	**M1. Precision:** Measures how accurately the system recommends relevant content, computed separately for male and female artists to assess group-wise performance disparities. **M2. F1 Score:** Harmonic mean of precision and recall; reflects the balance between false positives and false negatives in classification. **M3. Demographic Parity of Exposure:** Measures whether different groups receive equal exposure in recommendations, adjusted for position bias.
Are positive outcomes and errors equitably distributed across different demographic groups?	**Distribution of outcomes.** Examples: **A1.** Distribution of true positive rate **A2.** Distribution of utility per recommendation across different groups	**Disparate distribution of outcomes across groups.** Examples: **H1.** There is a significant difference in the distribution of true positive rates across groups. H2. There is a disparity in the expected utility of recommended items for different groups, leading to allocative unfairness in system outcomes	**M1. Recall Gap:** Measures the difference in true positive rates (recall) across groups; defined as the maximum gap in recall between any two protected groups. **M2. Dynamic Parity of Utility:** Measures whether different groups receive equal expected utility from recommendations, accounting for position bias and ranking policy over time.
	Distribution of errors. Examples: **A1.** Error rate in the output **A2.** Distribution of erroneous outcomes	**Disparate distribution of errors across groups.** Examples: **H1.** The error rate in the output is higher for different protected groups, indicating unequal performance across racial groups **H2.** The distribution of erroneous outputs differs across protected groups, leading to unequal error	**M1. Word Error Rate (WER):** Measures the accuracy of ASR systems by comparing machine-generated transcriptions to human-generated ground truth; calculated as the sum of substitutions, deletions, and insertions divided by the total number of words. **M2. Error Rate Equality Difference:** Measures the variation in false positive and false negative rates

		rates and potential harm	across identity terms; larger differences indicate greater unintended bias and deviation from equality of odds.
Does the AI system deliver outputs that are useful, inclusive, and satisfying to diverse users?	**User satisfaction with the AI system output.** Examples: **A1.** User satisfaction with the recommended package **A2.** User satisfaction interpreted through behavioral engagement with recommended content	**The user is not satisfied with the AI system's output.** Examples: **H1.** User satisfaction is disparate across group members, leading to imbalanced or unfair group recommendations **H2.** The user does not listen to many of the recommended tracks, indicating a misalignment between recommendations and user interest	**M1. Balance Error (BE):** Measures how far the recommended packages deviate from ideal fairness, where all users in a group are equally satisfied with the package based on their preferences for items and categories. **M2. User Satisfaction:** Measured as the number of tracks a user listens to from a recommended set; higher values indicate greater satisfaction.
	Alignment of AI system output with user desires. Examples: **A1.** Alignment between system output and user behavior conditioned on sociodemographics **A2.** Representation of an individual user within an instance or set of instances	**The AI system output is not aligned or relevant to the user's desires.** Examples: **H1.** Recommended content skews away from users' past preferences, reinforcing gender imbalance and reducing user autonomy **H2.** The inclusion score is low, indicating poor alignment between a user and the options relevant to them in an instance or set	**M1. Hellinger Distance:** Quantifies the divergence between the gender distribution of recommended content and the user's original listening history. **M2. Instance Inclusion:** Measures how well an individual is represented in the returned instance or set. Greater inclusion indicates better alignment between a user and the options relevant to them in the instance or set.

Interpersonal Harms

Table D4: Summary of the categories, attributes, and hazards extracted within interpersonal harms, with examples explained by various measures.

Category	Attribute	Hazard	Measure Examples
How well does the AI system design preserve privacy?	**The identifiability based on dataset composition.** Examples: **A1.** Identifiability of the entries of a dataset from obfuscated data **A2.** Identifiability of data entries	**High identifiability of individuals due to distinctive or revealing patterns in the dataset.** Examples: **H1**. A lot of information could be learnt about an individual in the dataset based on the publicly released obfuscated data. **H2.** Records in the dataset are distinguishable and easily identifiable	**M1. Inverse of the Trace of the Fisher Information Matrix:** Quantifies how difficult it is for an adversary to infer or reconstruct sensitive data entries from released data. **M2. K-Value:** Indicates the minimum group size required to make identification difficult.
	The implemented level of privacy protection for the dataset and the model. Examples: **A1.** Privacy of noisy feature extractor- How well the transformed feature hides the sensitive class B**2.** Privacy loss - how much information about an individual could leak through the outputs or parameters of the trained model.	**The implemented privacy protection is not enough OR too much.** Examples: **H1.** The added noise is not enough for protecting the privacy OR the added noise is too much for useful data analysis. **H2.** Differential privacy guarantees fail, allowing sensitive personal information (e.g., behaviors, identities) to be inferred from the model's decisions or shared parameters.	**M1. Privacy (z_i):** Evaluates privacy by checking how well a sensitive attribute (e.g., gender) is hidden in a noisy feature vector. It ranks the true sensitive class among all possible classes based on likelihood, and divides that rank by the total number of classes. **M2. (ε, δ)-Differential Privacy:** Quantifies how much information about any single individual's data can be inferred from a model's output. The parameters: ε (epsilon): The privacy loss- smaller values mean stronger privacy (outputs on similar datasets are nearly indistinguishable). δ (delta): the failure probability: The chance that the privacy guarantee does not strictly hold.
How vulnerable is an AI system to privacy attacks, and how much can be revealed?	**Likelihood of success for privacy attacks.** Examples: **A1.** Model's susceptibility to membership inference attack - how easily an attacker can determine whether a data point was used during model training.	**The system is highly susceptible to privacy attacks.** Examples: **H1.** The adversarial distance for a given input exceeds a threshold, revealing whether it was seen during training and thus leaking private information. **H2.** High prediction accuracy reveals insufficient	**M1. Adversarial Distance Score ($\hat{d}$):** Measures how much perturbation is required to turn a sample into an adversarial example. **M2. MAP (Maximum A Posteriori) Adversary Accuracy:** The adversary's probability of correctly guessing the private label based on the privatized data. Lower values indicate stronger privacy.

	A2. Success rate of adversarial inference	privacy protection	
	Information gained in case of a privacy attack. Examples: **A1.** Information leakage from private user data to the learner **A2.** Amount of information an adversary gains about the graph	**A high level of information is gained from the privacy attack.** Examples: **H1.** The learner or an adversary can infer sensitive user information if noise is insufficient or improperly calibrated. **H2.** The adversary gains significant information about graph structure or node identities, leading to potential re-identification of individuals.	**M1. Mutual Information:** Quantifies the amount of information leaked from a user's private data to the learning system in an online gradient descent algorithm. **M2. Amount of Leaked Information:** Measures how many nodes in the graph the adversary successfully re-identifies after a de-anonymization attempt. It reflects the amount of sensitive information exposed in the graph
To what extent does an AI system influence user behavior and trust, affecting autonomy?	**AI system's influence on the user.** Examples: **A1.** Behavioral trust **A2.** User trust in AI/ML output **A3.** Difference in item popularity preference vs. recommendation	**AI system influences the user behavior too much or too little.** Examples: **H1.** The number of times that users accept suggestions is significantly different from the number of times the system provides correct suggestions **H2.** Cutoff difference is different than the optimal value, implying users under or over-trust the system **H3.** Models continue to show popular items despite a user's interest to see niche/long-tail items	**M1. Acceptance of AI Suggestion:** The number of times a participant accepts the AI system's suggestions **M2. Difference of Decision Cut-off Points:** This quantifies how much the human decision threshold shifts when aided by the AI system versus unaided decision-making. **M3. Delta Group Average Popularity (ΔGAP):** Measures the shift in average popularity between items rated by a group and items recommended to that group. A value of 0 indicates the recommendations fairly reflect the group's interest in popular or niche content.
	User self-reported trust in the system. Examples: **A1.** Users' perceived trust in the information **A2.** Trust in the system — operationalized as confidence and perceived dependability of AI system decisions.	**Miscalibrated trust in the system.** Examples: **H1.** The user over- or undertrusts the information provided. **H2.** Misplaced reliance or rejection of AI-generated decisions	**M1. Assessed Trust in the Information Provided By the Articles:** Short news articles evaluated for relevance and credibility using Sundar's framework (1999) on a 5-point Likert scale. **M2. Overall Trust:** Derived from adapted Merritt scale items such as "I trust the tool," "I have confidence in the advice," and "I can depend on the tool."

How do users perceive their own and the AI's autonomy?	**Users' perception of their own autonomy.** Examples: **A1.** Perceived control - User's perceived ability to influence outcomes in the hiring process	**User has a misplaced perception of their own autonomy.** Examples: **H1.** Users do not have the opportunity to express opinions or influence processes	**M1. Perceived Control:** Four items from Saks and Ashforth (1999) measure how much applicants feel they can influence or control the outcome of the hiring process. They assess perceived autonomy, including whether success feels dependent on their actions versus external or uncontrollable factors.
	User's perception of the AI system's autonomy. Examples: **A1.** Perceived agency – the degree to which users believe the AI has autonomous decision-making capabilities.	**User has a misplaced perception of the system's autonomy.** Examples: **H1.** Overreliance or misattributed responsibility – if users attribute too much autonomy to the AI, they may over-trust it or fail to question its decisions, potentially leading to loss of user agency or accountability issues.	**M1. Mind Perception Agency:** Assesses the extent to which users believe an AI system can act independently, using five items related to self-control, moral reasoning, memory, empathy, and goal-directed behavior.
How do the outputs and behaviors of an AI system—and users' perceptions of them—impact users' well-being?	**Harmfulness of the content generated.** Examples: **A1.** Presence of hurtful words	**AI system's output is harmful.** Examples: **H1.** The model outputs hurtful words	**M1. The HONEST Score:** Measures the average proportion of hurtful completions generated by a language model when filling a set of sentence templates
	Perceived quality of the AI system. Examples: **A1.** Perceived helpfulness in diagnosis **A2.** Perceived risk	**Miscalibrated perception of the AI system's quality.** Examples: **H1.** Low level of helpfulness - at times, a high level of helpfulness might also lead to inappropriate overeliance. **H2.** Miscalibrated perceived risk of relying on AI systems - users may overestimate dangers, leading to distrust or rejection of beneficial technologies, or underestimate them, leading to overreliance and exposure to harm	**M1. Perceived Helpfulness:** Participants answer the Likert scale question: "[Version X] helped me think through the diagnosis and organize my thoughts.", rated on a 7-point Likert scale. **M2. Perceived Risk:** Participants rated the risk of various AI applications on a 0–10 scale after reading a related news story

	Perceived respectfulness of the AI system. Examples: **A1.** Perceived respectfulness of AI/ML process	**Miscalibrated perception of respectfulness of the AI system.** Examples: **H1.** People affected by AI/ML decision do not feel their experience is respectful and dignified	**M1. The Measure of Dehumanization in Formosa (2022):** Uses Bastian and Haslam's 10-item scale, which asks participants to rate how much they feel treated as less than fully human by a decision-maker (human or AI) in healthcare scenarios. It captures two dimensions: denial of human nature (e.g., being treated as emotionless or an object) and denial of human uniqueness (e.g., being seen as unintelligent or immature).

Social System Harms

Table D5: Summary of the categories, attributes, and hazards extracted within social system harms, with examples explained by various measures.

Category	Attribute	Hazard	Measure Examples
How well can stakeholders understand the AI system's behavior and output?	**Interpretability and understandability of the AI system.** Examples: **A1.** Complexity of individual model operations **A2.** User's perceived understanding	**The AI system is too difficult to interpret and understand.** Examples: **H1.** The model architecture is too complex, which makes it less interpretable. **H2.** The user perceives that they understand the model and its outputs well, when in reality, they have a poor understanding	**M1. Sentence Complexity (COMPLEX):** Calculated as the sum of complexity factors for each primitive (e.g., mathematical operations) used in the model. Higher values indicate more complex and less interpretable models. **M2. Perceived Understanding of the Model:** Calculated as the average self-reported agreement (on a 7-point Likert scale) with statements about understanding how the model works and the ability to predict its behavior.
	Quality of the post-hoc method. Examples: **A1**. Quality of post-hoc explanations of black-box models **A2.** Completeness of an explanation	**The quality of the post-hoc explanation is inadequate.** Examples: **H1.** Masked nodes do not decrease the quality of the predictions, indicating that the explanation model may not have correctly identified the truly influential nodes, leading to misleading or low-fidelity explanations. **H2.** The fraction of information in a subset is too low compared to a complete explanation, leading to incomplete or misleading explanations.	**M1. Fidelity Score:** Calculated as the drop in model accuracy when the most relevant nodes (identified by the explanation model) are masked. A higher drop indicates a more faithful explanation. **M2. Explanation Completeness:** The fraction of the total explanation (based on SHAP values) captured by a subset of features contributing to the model's output.
How does the resource intensity of AI system development contribute to environmental harm?	**Energy consumption of the process for creating an AI system.** Examples: **A1.** Energy consumption of neural networks **A2.** Electrical energy per sample processed	**The energy consumption is too high.** Examples: **H1**. Energy consumption of the chosen neural network is high **H2.** The accelerator uses a significant amount of electrical energy to process the samples	**M1. Average Energy Consumption:** Estimated based on multiply-and-accumulate (MAC) operations using computational load and platform-specific parameters. **M2. Electrical Energy Cost Per Sample Processed:** Calculated as Energy (J/sample) = Power (W) / Performance (Samples/s),

			using real-time power sampling tools during model training.
	Carbon emissions from the process of creating an AI system. Examples: **A1.** Carbon footprint of the total training protocol **A2.** Carbon emission of model training	**The carbon emissions are too high.** Examples: **H1.** The specific learning algorithm and training protocol at a given geographical location have a high carbon footprint **H2.** The model training results in a high CO2eq	**M1. Estimated GHG Emissions:** Calculated by multiplying the energy used for computing and communication by the carbon intensity of electricity in each device's region, reflecting how much CO_2 is emitted per kilowatt hour of local energy use. **M2. Amount of CO_2eq Produced:** Calculated using the ML Emissions Calculator, based on training time, GPU type, geographic location, and local carbon intensity of electricity.