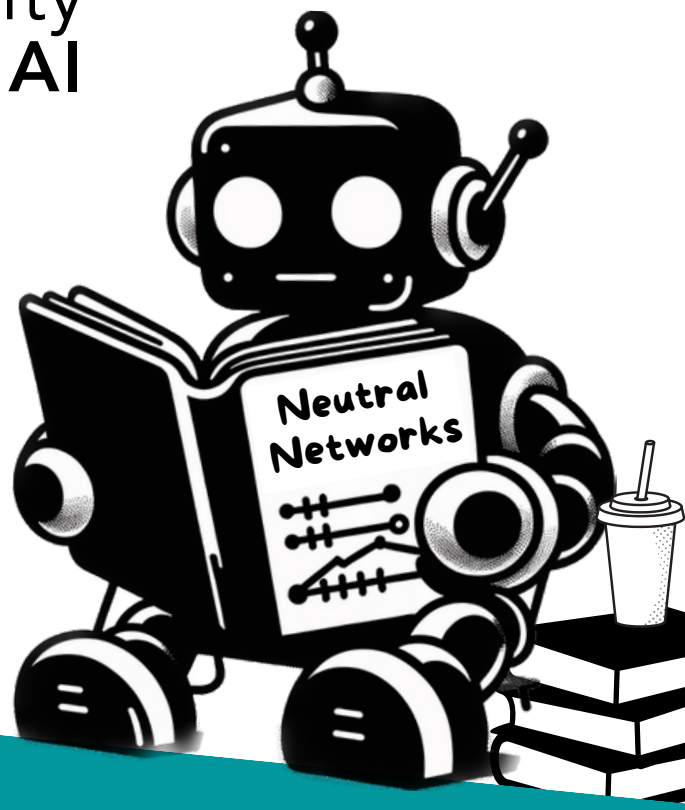
LEVELLING UP TO

AI ETHICS IN PRACTICE

Understanding paths to parity
and **FAIRNESS in** (*and with*) **AI**

A.I. DE-STATEMENTS
PRIVACY



FAIRNESS METRICS

Fairness metrics are the rules and standards designed to check if a given system is fair to everyone, regardless of their characteristics or background.

Fairness metrics in AI systems are used to evaluate, identify, and eliminate bias or discrimination.

TRANSPARENCY AND EXPLAINABILITY are essential for fair AI practices.

Biased information yields biased and innacurate decisions and perpetuate inaccuracy, unfair practices, and inequality. With AI poised to aid in decision making across many fields and disciplines, **FAIRNESS METRICS** are critical to creating and evaluating accurate and equitable decision making processes.

STATEMENTS

Shoshana Rosenberg

GET TO KNOW THEM
ASK FOR THEM BY NAME

BIAS

Bias, as you know, refers to an inclination or prejudice for or against one person or group, especially in a way that would be unfair.

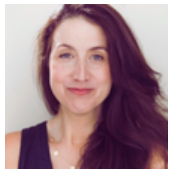
In AI, bias typically refers to systematic errors that create unfair outcomes, such as favoring one category of users over others. This can arise from various sources like biased training data, biased algorithms, or biased interpretation of results.

DATA BIAS: Arises from unrepresentative, biased, or incomplete data.

ALGORITHMIC BIAS: Occurs when algorithms process data in a way that reinforces stereotypes and/or unfair outcomes.

INTERPRETATION BIAS: When AI model results are interpreted or used in a biased manner *by humans*.

STATEMENTS



Shoshana Rosenberg

"A BIAS RECOGNIZED IS A BIAS STERILIZED." — BENJAMIN HAYDON

GROUP
INDIVIDUAL
SUBGROUP
CAUSAL
PREFERENCE -BASED

FAIRNESS

It is worth keeping in mind that fairness in AI is complex, in part because fairness itself is both complex and context-dependent, involving a larger framework of ethical, societal, and legal considerations.

As you will note, there are tradeoffs between different types of fairness metrics; and using multiple fairness metrics is often necessary to help ensure different dimensions of fairness. Assessments need to not only be multi-dimensional, but also dynamic- with continuous monitoring and adjustment.

"FAIRNESS IS WHAT JUSTICE REALLY IS."- POTTER STEWART

STATEMENTS

Shoshana Rosenberg

EXPLAINABILITY

Clear explanations and transparency are the key to building trust and ensuring accountability in AI, in DEI, and in Privacy.

EXPLAINABILITY METRICS are quantitative measures that work to assess how understandable an AI model's decisions are to humans. Where humans have clarity around the way the decisions are made by an AI, they are able to better and manage AI systems.

For human operators using AI to make decisions, explainability metrics provide insights that facilitate human decisions as to when to accept, reject, or override an AI system's recommendations.

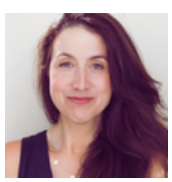
Shoshana Rosenberg

BREAKING IT DOWN

EACH TYPE OF FAIRNESS METRIC TAKES AIM AT EVALUATING AND PRESERVING SOMETHING DIFFERENT WITHIN AN AI.

Each type of fairness metric is also something HUMANS CAN USE AI TO ASSESS AND EVALUATE FAIRNESS IN OUR NON-AI SYSTEMS AND PROCESSES

Each fairness metric TIES IN WITH ONE OR MORE PRIVACY PRINCIPLES



Shoshana Rosenberg

GROUP FAIRNESS

DEMOGRAPHIC PARITY

WHAT DEMOGRAPHIC PARITY CHECKS FOR:

UNFAIRNESS ACROSS GROUPS

This metric ensures that the decision rate (loan approval, for example) is independent of sensitive attributes such as race and gender.

OF NOTE:

Demographic Parity ensures equal outcomes across groups but can ignore merit-based factors, potentially leading to underqualified candidates being favored.

USING AI TO FACILITATE DEMOGRAPHIC PARITY:

AI can be used to assess existing hiring or lending practices for disproportionate impacts on certain groups.

TRANSPARENCY AND ACCOUNTABILITY:

Ensuring Demographic Parity in AI requires clear documentation and justification of how demographic factors are considered (or not considered).

DATA MINIMIZATION AND PURPOSE LIMITATION:

It is critical that no more data is collected or used than is necessary and essential for the specific purpose.

STATEMENTS

Shoshana Rosenberg

GROUP FAIRNESS

EQUALITY OF OPPORTUNITY

WHAT EQUALITY OF OPPORTUNITY PARITY CHECKS FOR:

FAIRNESS FOR THE SIMILARLY QUALIFIED

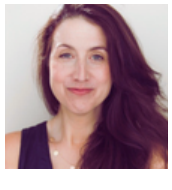
This metric focuses on fairness across positive results for individuals with similar qualifications. For example- do equally qualified candidates have the same chance of getting an interview, regardless of their background?

OF NOTE: This metric doesn't in any way address pre-existing inequalities that might have impacted a person's opportunities to attain certain qualification, and it cannot impact or guarantee equality of outcomes.

USING AI TO FACILITATE EQUALITY OF OPPORTUNITY:

AI can be used to help monitor outcomes to ensure organizational equal opportunities in scenarios like promotions or hiring.

STATEMENTS



Shoshana Rosenberg

TRANSPARENCY:

Equality of Opportunity Fairness Metrics mandate transparent decision making processes.

GROUP FAIRNESS

PREDICTIVE PARITY

WHAT PREDICTIVE PARITY CHECKS FOR:

PREDICTIVE ACCURACY ACROSS DIFFERENT GROUPS
Ensures that a prediction made by an AI system is equally accurate across demographic groups.

OF NOTE: Can be at odds with demographic parity or equal opportunity. *If the training data is not representative of the broader population, predictive parity will not protect against biased outcomes against underrepresented groups.*

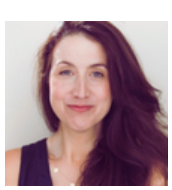
USING AI TO FACILITATE PREDICTIVE PARITY:

AI can be used to assess accuracy of diagnosis and prescribed treatments across groups of patients.

FAIRNESS:

Predictive parity relates directly to fairness in data analysis

STATEMENTS



Shoshana Rosenberg

GROUP FAIRNESS

EQUALIZED ODDS

WHAT EQUALIZED ODDS CHECKS FOR:

This metric evaluates the false negative and false positive rates- essentially where an incorrect result was provided- and tasks the AI with ensuring that both these error rates are similar across all groups.

OF NOTE: It does not account for the context or consequences of errors, which can vary greatly. This metric may force the model away from decisions based on the data to increase the error rate for one group to match an error rate for another group.

USING AI TO FACILITATE EQUALIZED ODDS:

AI can run automated D&I impact assessments through simulations to test how changes in processes or policies or criteria would impact different groups, ensuring that any changes made do not inadvertently introduce new biases.

FAIRNESS:

This metric works toward fairness and fair processing by striving to ensure that AI systems do not disproportionately misclassify individuals based on characteristics such as race or gender or age.

STATEMENTS

Shoshana Rosenberg

GROUP FAIRNESS

CONDITIONAL USE ACCURACY EQUALITY

WHAT CONDITIONAL USE ACCURACY EQUALITY CHECKS FOR: This metric aims to ensure that accuracy is consistent across different groups within the conditions of use and relies on the acccuracy and fairness of the underlying risk assessment model. *(This fairness metric seeks balanced accuracy given the same conditions- as opposed to Equalized Odds, which looks to balance error rates overall.)*

OF NOTE: This metric relies upon having comprehensive and representative data for all groups to give accurate fairness assessments. It is hard to implement because in many scenarios related to humans, measuring "similar conditions"can prove challenging, as can designing the metric to account for contextual nuances. ***Much like with Equalized Odds, the risk of overcorrecting for one group at the expense of others is significant.***

USING AI TO EVALUATE CONDITIONAL USE ACCURACY EQUALITY:
AI could be used to make an analogous evaluation of performance reviews across demographic groups and whether they receive similar outcomes (pay raises, promotions, etc.)

NON-DISCIMINATION:
This fairness metric, in aiming for similar accuracy for all groups at each level of risk or decision threshold, works to support non-discrimination.

STATEMENTS

Shoshana Rosenberg

GROUP FAIRNESS

EQUAL OPPORTUNITY

WHAT EQUAL OPPORTUNITY FAIRNESS CHECKS FOR:

This metric focuses on ensuring that qualified members of all groups should have an equal chance of being correctly identified for a given outcome or benefit.

OF NOTE: Again, the effectiveness of this fairness metric can vary based on the context and nature of the AI application, and by addressing one aspect of fairness, does not account for complexities such as the differing consequences that a false negative (failure to be properly identified for an outcome or benefit) could have on members of different groups.

USING AI TO FACILITATE EQUAL OPPORTUNITY:

AI could be used to analyze patient records and treatment data to assess whether all demographic groups with the same medical conditions have equal access to high-quality care and to investigate inconsistent applications of essential treatments or preventative care.

NON-DISCRIMINATION:

Equal Opportunity as a fairness metric supports the principle of non-discrimination by striving for equitable outcomes across all groups.

STATEMENTS

Shoshana Rosenberg

GROUP FAIRNESS

DISPARATE IMPACT

WHAT DISPARATE IMPACT CHECKS FOR:

This metric assesses the impact of an AI system's decisions on different groups, aiming to identify biases.

OF NOTE: Disparate impact can potentially identify legal and ethical issues as it may indicate/identify discrimination, eveni if intentional. When balancing a model, care must be taken not to overcorrect and create reverse discrimination.

USING AI TO ADDRESS DISPARATE IMPACT:

This is another situation in which healthcare policies and practices can be improved by AI aiding the statistical analysis of outcomes and resource distribution.

FAIRNESS:

This aims to identify and help remedy disparate impacts.

STATEMENTS

Shoshana Rosenberg

CAUSAL FAIRNESS

COUNTERFACTUAL FAIRNESS

WHAT COUNTERFACTUAL FAIRNESS CHECKS FOR:

This metric checks to see if counterfactual fairness has been achieved- which is to say, it tests the system to see if a decision would remain the same if a sensitive attribute for a given individual were changed (a counterfactual piece of data.)

OF NOTE: This requires complex modeling to test hypothetical scenarios and comprehensive data to include all relevant factors that should be tested.

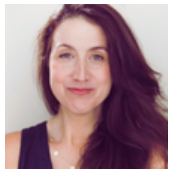
USING AI TO EVALUATE COUNTERFACTUAL FAIRNESS:

AI can be used to assess whether educational policies or procedures would provide the same opportunities to students if their socioeconomic status or other sensitive attributes were different.

TRASNPARENCY:

The rational behind all decisions should be justifiable and understandable, and independent of an individual's characteristics.

STATEMENTS



Shoshana Rosenberg

INDIVIDUAL FAIRNESS

FAIRNESS THROUGH AWARENESS

WHAT FAIRNESS THROUGH AWARENESS CHECKS FOR:

This focuses on treating individuals fairly by considering their unique characteristics and contexts. The metric advocates for decision-making processes that take into account individual nances, not relying solely on generalized group attributes.

OF NOTE:

This is a complex and data intensive endeavor, and a lack of very detailed information for all individuals will lead to challenges.

USING AI TO EVALUATE FAIRNESS THROUGH AWARENESS:

AI can be used to analyze data of students to evaluate whether educational programs at an institution address the specific needs and backgrounds of its students.

STATEMENTS

Shoshana Rosenberg

PRIVACY BY DESIGN:

The large amount of data required highlights the need for Privacy by Design means of using so much personal data.

SUBGROUP FAIRNESS

FAIRNESS UNDER COMPOSITION

WHAT FAIRNESS UNDER COMPOSITION CHECKS FOR:

This metric is focused on the recognition that individuals often belong to multiple demographic subgroups, and works to address fairness as it applies to the resulting intersectionality.

OF NOTE: This is a highly complex undertaking requiring comprehensive data in order to protect the specific aspect of fairness that it targets.

USING AI TO EVALUATE FAIRNESS UNDER COMPOSITION:

AI could be used to discern whether diversity initiatives give equal support to employees or students who belong to multiple minority groups. If by privacy-by-design diversity data and employee feedback isolation, AI could safely be used to analyze employee feedback, rates of promotion and participation to evaluate impact where there is intersectionality.

STATEMENTS

Shoshana Rosenberg

FAIRNESS:

Striving to ensure fairness evaluation is applied for individuals belonging to multiple sensitive categories.

PREFERENCE-BASED FAIRNESS

WHAT PREFERENCE -BASED METRICS CHECK FOR:

Preference-based fairness looks to respect individual preferences in decision making processes- for a more personalized fairness. *(Think of grocery replacement offerings where you place an order and something is out of stock. The equivalent offering from a stock and price vantage point may run against your personal aversions, preferences, allergies...)*

OF NOTE: Accurately capturing individual preferences requires robust mechanisms, a great deal of data. Weighing distinct subjective notions of fairness is liable to result in data that conflicts.

USING AI TO EVALUATE PREFERENCE-BASED FAIRNESS:

AI can be used to assess whether workplace remote work and other policies are able to align with the diverse preferences of employees based on analysis of requested feedback.

STATEMENTS

Shoshana Rosenberg

THIS ONE IS A BITOF A STRETCH TO ALIGN WITH PRIVACY PRINCIPLES: but it does respect the individual and the importance of subjective fairness.

FAIR REPRESENTATION LEARNING

WHAT FAIR REPRESENTATION LEARNING IS FOCUSED ON:

Fair representation learning, while not a metric, is key because it involves creating data representations that do not PERPETUATE BIASES against groups.

OF NOTE: Ensuring that learned representations are truly unbiased can be a challenge, as biases can be deeply embedded in data.

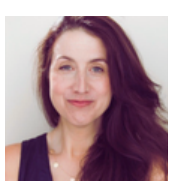
USING AI TO EVALUATE FAIR REPRESENTATION:

AI can be used to help plan and design job applicant and evaluation processes to focus solely on factors directly relevant to job performance, removing or anonymizing irrelevant personal informatio

NON-DISCRIMINATION:

This approach directly supports non-discrimination in the processing of personal data.

STATEMENTS



Shoshana Rosenberg

NEW PRIVACY METRICS THAT SHOULD BE SET IN PLACE

Privacy is critical to fairness.

Fairness metrics should also exist specifically to ensure that AI systems respect and protect the personal data and data rights of individuals. I propose the creation of the following Privacy-Centric Fairness Metrics to promote lawfulness, transparency, and trust.

Data Exposure Limitation Metric:

This metric would measure the extent to which personal data is exposed to the AI during the decision making process, ensuring that only the essential elements are used for the given task. This could track and score the data minimization practices within a system.

Consent Fidelity Score:

This metric could assess how closely the data usage aligns with user consent, evaluating whether the AI system used personal data strictly within the bounds of what users have agreed to.

Data Retention/Purpose Alignment Metric:

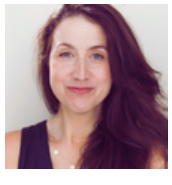
This metric, though likely a nuisance to set up for most systems, could be developed to measure the alignment between the stated purpose for data collection, the requirements for retention and use that accompany that purpose, and how long the data is actually retained and used.

Equitable Data Access and Correction Metric:

This data would measure what I call the "HOV lane" for the system to easily allow individuals needed access and the ability to correct or delete their data, assessing responsiveness of the AI system to DSARs.

Transparent Data Pathway Metric:

This metric could be created to assess the transparency of the data lifecyles within an AI system, from collection to processing and through to deletion.



Shoshana Rosenberg

A.DE STATEMENTS PRIVACY

DO YOU WANT TO PLAY A GAME?

Stay tuned. I hope to have one for you: "Promptly".

STATEMENTS

Shoshana Rosenberg

save for later →

PRIVACY STATEMENTS