Ivey | Publishing

W43809

Teaching Note

MACHINE LEARNING BIAS: ALGORITHMS IN THE COURTROOM[1]

Professors Yasser Rahrovani and Lauren Cipriano wrote this teaching note as an aid to instructors in the classroom use of the case Machine Learning Bias: Algorithms in the Courtroom, no. W42647. This teaching note should not be used in any way that would prejudice the future use of the case.

SYNOPSIS

This case examines the complexities of using machine learning (ML) and artificial intelligence (AI) in judicial decision-making through the development and application of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) tool for assessing recidivism risk. The case contrasts two perspectives of fairness: one highlighting algorithmic fairness in outcome and the other in procedure. Students explore the ML development cycle, identifying nine types of biases and their sources. The case fosters understanding of the ML development cycle, bias identification, and management responsibilities to navigate artificial intelligence (AI)'s societal and ethical challenges responsibly.

LEARNING OBJECTIVES

By working through the case and assignment questions, students will be able to

- understand the machine learning development cycle;
- identify and categorize nine types of biases emerging in AI/ML development;
- classify machine learning biases in three reinforcing categories of data, algorithms, and user interactions;
- map biases across the stages of the machine learning development cycle to understand their origins and impact; and
- analyze competing definitions of fairness and their practical implications.

POSITION IN COURSE

The case is suitable for undergraduate- and graduate-level business classes as well as executive education, particularly in core information technology management, innovation management, fundamentals of AI, and governance courses.

[1] All case facts and case research that are referenced in this teaching note are supported by the sources and information contained in the associated case document (product No. W42647) unless otherwise specified.

RELEVANT READINGS

Artificial Intelligence Bias

- Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan, "A Survey on Bias and Fairness in Machine Learning," Cornell University, August 23, 2019. https://arxiv.org/abs/1908.09635v3.
- Harini Suresh and John Guttag, "A Framework for Understanding Sources of Harm Throughout the Machine Learning Life Cycle," *Equity and Access in Algorithms, Mechanisms, and Optimization*, EEAMO '21: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 1–9. https://doi.org/10.1145/3465416.3483305.

COMPAS

- William Dieterich, Christina Mendoza, and Tim Brennan, "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity," Northpointe, July 8, 2016, https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.
- Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias," ProPublica, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
- Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin, "How We Analyzed the COMPAS Recidivism Algorithm," ProPublica, May 23, 2016, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.
- Dana Casadei, "Predicting Prison Terms and Parole," Downtown Newsmagazine, March 24, 2020, https://www.downtownpublications.com/single-post/2020/03/24/predicting-prison-terms-and-parole.
- Julia Dressel and Hany Farid, "The Accuracy, Fairness, and Limits of Predicting Recidivism," *Science Advances* 4, no. 1 (2018). https://doi.org/10.1126/sciadv.aao5580.
- Sam Corbett-Davies, Emma Pierson, Avi Feller, and Sharad Goel, "A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear," *The Washington Post*, October 17, 2016, https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/
- Eugenie Jackson and Christina Mendoza, "Setting the Record Straight: What the COMPAS Core Risk and Need Assessment Is and Is Not," Harvard Data Science Review, March 31, 2020, https://doi.org/10.1162/99608f92.1b3dadaa.
- Julia Angwin and Jeff Larson, "ProPublica Responds to Company's Critique of Machine Bias Story," ProPublica, July 29, 2016, https://www.propublica.org/article/propublica-responds-to-companys-critique-of-machine-bias-story.
- Anthony W. Flores, Kristin Bechtel, and Christopher T. Lowenkamp, "False Positives, False Negatives, and False Analyses: A Rejoinder to 'Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks,'" *Federal Probation* 80, no. 2 (2016): 38–42. https://www.uscourts.gov/sites/default/files/80_2_6_0.pdf.
- Anna Maria Barry-Jester, Ben Casselman, and Dana Goldstein, "The New Science of Sentencing," The Marshall Project, April 8, 2015, https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing.

SUPPLEMENTAL MATERIALS

Pre-class Task

- Yasser Rahrovani, “ML Bias: AI in the Courtroom,” January 21, 2025, in *CaseCast*, podcast, MP3 audio, 42:56, https://www.buzzsprout.com/2293124/episodes/16470939.

Bias

- CNBC, “How AI Could Reinforce Biases in the Criminal Justice System,” March 18, 2019, YouTube video, 0:19, https://youtu.be/ZMsSc_utZ40?feature=shared.
- ITiCSE 2020, “411 Auditing the COMPAS Recidivism Risk Assessment Tool Predictive Modelling and Algorithmic Fairn,” September 2, 2020, YouTube video, 4:35, https://youtu.be/HBpQbmwbcCo?feature=shared.
- Sky News, “Racial Bias in AI: Man Wrongly Identified by Facial Recognition Technology,” November 10, 2023, YouTube video, 1:23, https://youtu.be/Cx284WjpEQY?feature=shared.
- The Economist, “How to Make Computers Less Biased,” February 10, 2022, YouTube video, 8:53, https://youtu.be/IzvgEs1wPFQ?feature=shared.
- Yasser Rahrovani, “Fundamentals of AI for Managers,” April 26, 2024, in *CaseCast*, podcast, MP3 audio, 35:55, https://www.buzzsprout.com/2293124/episodes/14962190.

CONCEPTS AND MODELS

AI refers to the field of computer science focused on creating systems capable of performing tasks that typically require human intelligence. AI systems achieve this through techniques like ML, which is a field of statistical inference that identifies patterns in existing data sets and applies these patterns to make predictions on new, unseen data. ML algorithms are ubiquitous, influencing areas such as personalized recommendations on streaming platforms, facial recognition technology, virtual assistants, real-time transcription services, image tagging, and even critical decisions like job-application screening and medical approvals. At its core, ML involves discovering functional relationships in large data sets and using these relationships to infer outcomes in new scenarios.

Bias in Artificial Intelligence Systems

The increasing use of AI and ML systems in various aspects of life has made it essential to address the issue of bias within them. AI systems are increasingly used to make life-changing decisions in sensitive environments such as judicial courts, which underscores the need to ensure that these decisions do not discriminate against certain groups. ML algorithms, while beneficial, are susceptible to biases that can lead to unfair or discriminatory outcomes.

The creation of bias in AI systems is a complex issue arising from multiple sources. It is important to understand when and how bias might be introduced throughout ML development to prevent and mitigate undesirable consequences. Therefore, it is insufficient and shortsighted to simply say that data is biased for two reasons. First, data is one among many elements in an AI system. Others include user behaviours and design choices, such as algorithms. Second, data is not a static artifact but the result of a dynamic and evolving process influenced by other elements.[2] Understanding the implications of each stage in the data-generation process can lead to direct and meaningful ways to address harmful downstream consequences.

[2] Harini Suresh and John Guttag, “A Framework for Understanding Sources of Harm Throughout the Machine Learning Life Cycle,” *Equity and Access in Algorithms, Mechanisms, and Optimization*, EEAMO ’21: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 1–9. https://doi.org/10.1145/3465416.3483305. See Relevant Readings.

Machine Learning Development

The typical ML workflow involves several stages, each crucial to the system's performance.[3] These stages—which include data collection, data preparation, model development, evaluation, post-processing, and deployment—are interconnected and iterative. Below is a breakdown of each step, illustrating their roles in the ML life cycle. Each step contributes to the overall system and introduces opportunities for optimization and error. Understanding this process is vital for anticipating and mitigating potential sources of harm throughout the ML life cycle (see Exhibit TN-1).

Data Collection

The first step in the ML development process involves assembling a data set from a target population. This step entails defining the population of interest and selecting features and labels to represent relevant characteristics. As it is not feasible to capture data from the entire population, a representative sample is used. This step can involve leveraging existing data sets or creating new ones through specific collection methods.

In COMPAS's case, data collection involved gathering historical data from criminal justice systems, for example Florida. The target population consisted of individuals previously involved in the criminal justice system. The question is then what to measure or collect about the target population. Variables were selected from demographic and familial information, criminal history, and personal background questionnaires. The data included both direct records, such as the number of prior arrests and convictions, and survey-based responses that aimed to capture personal and social dynamics, including questions such as "Was one of your parents ever sent to jail or prison?" "How many of your friends/acquaintances are taking drugs illegally?" or "To what extent do you agree that a hungry person has a right to steal?"

Data Preparation

Once the data set has been collated, preprocessing is essential to clean and structure it for use. This may involve handling missing values, normalizing scales, or transforming features into usable formats. The data set is typically split into two subsets: (1) training data for building the model; and (2) test data to assess final performance and internal validation. For example, if the data set included responses from 1,000 individuals, 700 cases might have been allocated for training, 200 for internal validation during the training phase, and 100 for final testing.

These preprocessing steps ensure that the data set is structured and ready for model training. However, it is worth noting that choices made during this stage could have introduced biases, such as oversimplifying nuanced variables or inadvertently emphasized certain data patterns over others.

In COMPAS's development, data preparation could involve handling missing values, simplifying variables, and normalizing continuous measurements. For example, missing data, such as gaps in employment history or incomplete education records, might have been filled in using statistical methods like interpolation. Variables such as prior offences could have been grouped into broader categories—like distinguishing between misdemeanours and felonies—rather than using specific crime types. Continuous data, such as age at first arrest or income levels, may have been normalized to fit within a 0–1 scale.

[3] Suresh and Guttag, "A Framework for Understanding Sources."

Model Development

This stage involves training the algorithm using the prepared data to optimize for specific objectives, such as minimizing prediction errors. The development process often includes experimenting with different algorithms and optimization techniques. The best-performing algorithm is selected based on its ability to predict the output based on the inputs.

In developing the COMPAS model, the team first needed to select a specific algorithm, such as logistic regression or decision trees, and define the objective function, like minimizing prediction error in recidivism risk assessments. During training, the model aimed to learn a function that mapped inputs—such as demographic data, prior convictions, and responses to survey questions—to the output: whether an individual would be rearrested within two years.

The developers likely experimented with several model configurations, such as varying the algorithm's complexity or the weight assigned to different input variables. These configurations were compared based on their performance on labelled data, and the best-performing model was chosen. For instance, the team might have tried models that treated prior arrests and survey responses differently to understand their predictive power. Ultimately, the goal was to identify a configuration that accurately predicted recidivism while balancing interpretability and robustness.

Model Evaluation

Following development, the model's accuracy and robustness are assessed using the test data (internal validation). The test data remains untouched until this stage, ensuring that the model's performance reflects its ability to handle previously unseen information. Performance metrics are chosen based on the problem context and include measures such as accuracy, precision, recall, or error rates. This stage may also involve benchmarking against external data sets to externally validate the model and its inner algorithms' applicability across diverse scenarios, which is known as external validation.

For COMPAS, the evaluation process involved assessing the model's performance using test data that had been kept separate from training data to ensure unbiased results. The accuracy of the model was measured against its ability to predict recidivism outcomes correctly. Several performance metrics were considered—for example, judges might focus on the false negative rate (i.e., failing to identify individuals likely to reoffend), while developers might prioritize metrics like overall accuracy or the false positive rate (i.e., incorrectly labelling individuals as high-risk).

In addition, COMPAS could have been tested on external data sets, such as records from other jurisdictions or time periods, to evaluate its robustness. For example, while the model may have been optimized using data from Florida, it might have been further tested on data sets from Pennsylvania to confirm its generalizability across different populations and conditions. This step would ensure the model's outputs aligned with real-world applications, although disparities in error rates across demographic groups raised concerns about fairness and equity.

Model Post-processing

Depending on the desired output format, additional transformations may be applied to the model's results. For instance, a continuous output, such as a probability, might be converted into a binary or categorical classification for practical use.

In COMPAS, the post-processing stage involved transforming the model’s output—a continuous score indicating the likelihood of recidivism—into a format suitable for judicial use. For instance, the model’s raw output, a risk score ranging from 1 to 10, might have been categorized into descriptive levels such as low-risk, medium-risk, or high-risk to make it easier for judges and parole officers to interpret. Alternatively, the team could have opted for binary classifications, such as “eligible for parole” or “not eligible.”

This step required making critical decisions, such as selecting the thresholds that determine which scores fall into each category. For example, a score above 7 might be classified as high-risk, influencing decisions about incarceration or parole. These choices were pivotal because they directly affected how the model’s recommendations were perceived and applied in practice, potentially introducing further bias if thresholds were not carefully calibrated to avoid disproportionate impacts on specific demographic groups.

Model Deployment

Finally, the trained model is deployed in real-world applications. Deployment involves integrating the model into operational systems, creating interfaces for users, and establishing mechanisms to handle feedback. Real-world deployment often reveals disparities between the training population and the actual user base, necessitating continuous monitoring and adaptation.

To deploy COMPAS in real-world judicial systems, several practical considerations were addressed. The model needed to be integrated into courtroom decision-making processes, requiring the development of user interfaces that displayed the risk scores and recommendations clearly. For instance, judges might receive visualizations of a defendant’s risk level, supported by explanations or context to aid their understanding. It should be noted that while a model is developed using data from a specific context (e.g., data from Florida), it may be deployed to other contexts (e.g., in states like Wisconsin and New York), where demographic and social differences could impact its effectiveness. This raised concerns about how well the model was generalized to new contexts and how disparities might arise in its application. In addition, the use of an AI-based system can also vary as it is deployed in a user population, such as using COMPAS for resource allocation rather than sentencing.

Sources of Bias in Machine Learning Development

This case introduces three categories of ML biases, each including three subtypes.[4] To be able to address AI bias, we first need to understand how these different types of bias emerge, interact, and amplify each other.

1. *Data (to algorithm) bias*. Many AI systems are data-driven, and biases in the training data can lead to algorithms that reflect and amplify these biases. This includes
   a) measurement bias, where measures are selected or measured inaccurately;
   b) representation bias, where the sampled data under- or overrepresents population diversity; and
   c) aggregation bias, when a one-size-fits-all model is used for data that includes different underlying subgroups that should be treated differently.

2. *Algorithmic (to users) bias*. Even with unbiased data, algorithms can exhibit bias due to design choices or how they are optimized. This includes

[4] Suresh and Guttag, “A Framework for Understanding Sources”; Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan, “A Survey on Bias and Fairness in Machine Learning,” Cornell University, August 23, 2019, https://arxiv.org/abs/1908.09635v3. See Relevant Readings.

a) learning bias, when modelling choices amplify performance disparities;
b) user-interaction bias, which arises from how the AI tool interacts with users; and
c) evaluation bias, when the benchmarks and metrics used to assess a model's performance do not fully represent the deployment populations.

3. *Users (to data) bias*. User-related biases can also create a feedback loop, where user behaviour generates more biased data. This includes
   a) historical bias that reflects existing societal biases and inequalities used as training data;
   b) deployment bias, when there is a mismatch between how a model is intended to be used and how it is actually used; and
   c) self-selection bias, when the individuals who are included in a data set are not a random sample of the population but have selected themselves into the data.

It is important to recognize the circular relationship between the three categories of biases—data, algorithms, and users—which will be important when trying to mitigate bias. Addressing bias in AI requires a deep understanding of how biases are created, the originating sources, the interplay between data, algorithms and user biases, and the need to develop specific mitigation strategies relevant to the context in which the AI system is being used. In addition, given the trade-offs between types of biases and organizational priorities and values, it should be noted that fairness is not a one-size-fits-all concept.

This case and associated exercises will help students to articulate the nine types of biases within these three categories.

Data (to Algorithm) Bias

1. <u>Measurement Bias (in Variable and Data Selection)</u>

Measurement bias occurs when the features and labels used in the prediction problem are chosen, collected, or computed in a way that introduces systematic inaccuracies.

This bias arises from the use of proxies (i.e., features or labels that are proxies for an underlying construct that is difficult to measure directly, like a five-star rating for measuring satisfaction on Amazon.com, Inc.); from differential measurement (when the way a feature or label is measured varies across groups such as if one group is monitored more frequently, there will be more observations for that group, more skewed data); and from inaccurate proxies (when the proxies used to represent a particular construct differ in quality or accuracy across different groups).

COMPAS used "arrest" as a proxy for crime and "rearrest" as a proxy for recidivism. There is measurement bias because police practices and racial profiling can cause minority communities to have higher arrest rates without committing more crimes, so does not accurately reflect actual criminal behaviour. This means that these proxies are differentially measured, since there is a different relationship between crime and arrest for people from different communities. Any choice of variable can lead to certain measurement biases.

There is also measurement bias in the use of "rearrest" as a proxy for recidivism. This indirect measurement can introduce systematic inaccuracies, as rearrest rates can be influenced by factors unrelated to actual reoffending, such as racial profiling. This results in biased data that skews the model's predictions.

2. Representation Bias

This occurs when the development sample (i.e., the data used to train the model) under- or overrepresents some part of the population that the model is intended to serve. The focus is on whether the data accurately reflects the diversity of the target population. If the training data does not adequately and fully capture this, it leads to a model that does not generalize well for all groups. The question to ask in representation bias is: "Does the data set include a fair representation of all the relevant groups?"

This bias arises in a number of ways: when there is a non-representative sample (the data used to train a model does not include all relevant groups); the target population does not reflect the use population (the data is not representative of the context where the model will be deployed); and through implicit bias—even if a data set does not include explicit demographic information, it may contain other factors correlated with those demographics, leading to underrepresentation or misrepresentation of certain groups (e.g., even though race was not included as an explicit input, correlations between the data and race led to disparities).

COMPAS's data was less accurate for Black defendants, suggesting that the training data overrepresented their crime rate or inadequately captured their experiences. Black defendants are also overrepresented when using arrest records as a proxy for criminal behaviour, given policing practices and racial profiling. Another example is the fact that COMPAS uses questions about a defendant's friends' drug use or a parent's arrest, which correspond to socio-economic factors that disproportionately affect Black populations. So "although the data used by COMPAS do not include an individual's race, other aspects of the data may be correlated to race that can lead to racial disparities in the predictions."[5]

In May 2016, ProPublica published an analysis of the efficacy of COMPAS on more than 7,000 individuals arrested in Broward County, Florida, between 2013 and 2014. This analysis indicated that the predictions were unreliable and racially biased. COMPAS's overall accuracy for white defendants is 67.0 per cent, only slightly higher than its accuracy of 63.8 per cent for Black defendants.[6]

3. Aggregation Bias

This occurs when a one-size-fits-all model is used for data that includes different underlying groups or types of examples that should be treated differently. This means that the model assumes the relationship between inputs and outputs is consistent across all data subsets, which is often not the case. Specifically, it arises when subgroup differences are ignored and generalization is made about groups to individuals, assuming a consistent relationship that is not true.

The COMPAS algorithm might exhibit aggregation bias by applying the same risk prediction model to all defendants, without considering how factors like socio-economic status or community conditions might impact recidivism differently across racial or ethnic groups. For example, a history of prior arrests might have different implications for future recidivism risk depending on the community a person is from and their socio-economic status. The COMPAS model uses a one-size-fits-all approach without accounting for these differences between individuals, which does not optimally serve any specific group.

[5] Julia Dressel and Hany Farid, "The Accuracy, Fairness, and Limits of Predicting Recidivism," *Science Advances* 4, no. 1 (2018). https://doi.org/10.1126/sciadv.aao5580. See Relevant Readings.
[6] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias," ProPublica, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. See Relevant Readings.

Algorithm (to Users) Bias

4. Learning Bias

This arises when modelling choices amplify performance disparities across different examples or groups in the data. This means that the way the model learns from the data exacerbates existing biases, leading to unequal outcomes for different groups. This is the type of bias that is added by the algorithm. The algorithmic design choices, such as the use of certain optimization functions, regularizations, choices in applying regression models on the data as a whole or considering subgroups, and the general use of statistically biased estimators in algorithms, can all contribute to biased algorithmic decisions that can skew the outcome of the algorithms. The focus of learning bias is on the model's objective function and optimization process.

This type of bias arises in three ways: (a) when a model's objective function (what it is trying to optimize) can inadvertently damage another objectives; (b) when the model is trained only to maximize accuracy without considering fairness criteria; and (c) when there is an inherent trade-off between objectives (e.g., there may be trade-offs between different objectives like accuracy and fairness, meaning that optimizing one can compromise the other).

In the context of COMPAS, if the algorithm's objective is to maximize overall prediction accuracy, it might prioritize learning patterns that are common in the majority group, which may lead to the algorithm being less accurate for minority groups that have different patterns of behaviour. This can also manifest as a tendency to produce more false positives for certain groups. This occurs because, unless two groups have equal base rates, building an algorithm that produces well-calibrated scores as well as equal false positives is not possible. The model might learn to be more cautious (resulting in more false positives) when predicting recidivism for a group with a higher base rate of recidivism. Also, having a simple model for a complex system, leads to learning bias:

> Northpointe does not reveal the details of their COMPAS software […] we have shown that their prediction algorithm is equivalent to a simple linear classifier. In addition, despite the impressive collection of 137 features, it would appear that a linear classifier based on only two features—age and total number of previous convictions—is all that is required to yield the same prediction accuracy as COMPAS. This finding is supported by earlier work that showed how to create simple, transparent, and interpretable predictive rules for recidivism prediction. When applied to the same Broward County data set, the authors found that simple models afforded similar predictive accuracy as COMPAS.[7]

5. User-Interaction Bias

This arises from how a tool interacts with users, including the user interface and the presentation of information. The focus of user-interaction bias is on the interface between the user and the system, and whether that interface itself creates bias. It arises from a tool's user interface (presenting a risk score as a binary number has more polarized impact rather than a spectrum of 1 to 7); from ranking bias (the order in which options are presented can affect user choices); and from popularity bias (items that are already popular might be favoured by a system, leading to a feedback loop that amplifies the popularity of some items at the expense of others).

In the context of COMPAS, judges who are presented with a risk score might treat defendants with high-risk scores more harshly. This can happen even if the risk score is not a perfect predictor of recidivism. The

[7] Dressel and Farid, "The Accuracy, Fairness, and Limits of Predicting Recidivism."

way the risk scores are presented in the user interface may cause judges to overemphasize them in their decision-making. This means that a tool intended to reduce bias can inadvertently amplify it through how users interact with it. This can also happen when a judge uses the COMPAS tool to determine the level of supervision that a person receives when on parole. The user interface and the interaction of the users (judges' decisions) can be influenced by the tool. For example, the impact of what an algorithm shows to a judge regarding the risk of new violent criminal activity can differ when it is presented as a binary variable (yes or no) versus when as a scale of 1–7. The binary option can lead to harsher penalties.

6. Evaluation Bias

This occurs when the benchmarks and metrics used to assess a model's performance do not fully represent the diverse populations affected by it or when they are otherwise inappropriate for the evaluation. The focus of evaluation bias is on the evaluation process itself and the benchmarks that are chosen. It is about whether the metrics and data used to evaluate a model's performance accurately reflect the real-world impact and whether those metrics obscure or amplify the model's unfairness.

It arises from (a) inappropriate benchmarks (not reflecting the diversity of the population); (b) limited scope of metrics (focusing on a single metric, like overall accuracy, hiding other disparities like false positive rates); (c) overfitting to a benchmark (not performing well in real-world settings); or (d) a lack of external validation (not tested externally on new data sets).

In the case of COMPAS, the evaluation may have used a limited set of features or a specific data set that did not capture the full diversity of the population, leading to an algorithm that spits out biased outcomes. For example, if the model was externally evaluated based on accuracy for predicting recidivism for white defendants, that accuracy might not generalize well to Black defendants if there are different factors influencing recidivism for this group. Further, COMPAS was a closed software system, so analysts, judges, and other stakeholders could read the scores but could not analyze the underlying data and evaluation.

Users (to Data) Bias

7. Historical Bias

This reflects existing societal biases and inequalities that are embedded in historical data that is used to train algorithms. This means that even if data collection and measurement are perfect, the data can still reflect historical disparities and prejudices, which the algorithm will then learn. The focus of historical bias is on the pre-existing conditions and societal structures that have led to skewed data.

This type of bias arises from encoded societal norms (e.g., historical patterns of policing); past discrimination (historical disparities in treatment toward specific groups, such as race and gender); reinforcing stereotypes (harmful stereotypes against certain groups); and unequal opportunity (systematic disparities in education, housing, and employment).

The COMPAS algorithm is trained on historical crime data, which include arrest records. These records may be biased, for example, if certain groups are policed more heavily than others. Even if COMPAS uses unbiased scoring rules, the historical context of racial disparities in crime and policing leads to biased outcomes, reinforcing harmful stereotypes against certain groups. Because the model is trained on historical data, the predictions reflect historical forms of discrimination. As a result, COMPAS scores can discriminate if they mirror patterns in historical crime data.

8. Deployment (or Behavioural) Bias

This occurs when there is a mismatch between how a model is intended to be used and how it is used in practice. For example, deployment bias happens when a loan approval model is trained to predict risk, but it is used by human decision-makers who override or misuse its outputs. Another example is a hiring model trained on resumes is deployed in a chatbot that screens applicants, but the interaction context wasn't accounted for. The bias is introduced **during use**, **not training**, often due to changes in context, interpretation, or feedback loops.

Deployment bias arises through user interpretation (a judge may place too much weight on a risk score, or they may interpret it differently based on personal biases); system misuse (when used for a purpose other than what it was originally designed for); automation bias (user tendency to favour the recommendations of an automated system); or confirmation bias (users more likely to accept information that confirms their existing biases).

Judges varied in their use of COMPAS, which compromised the goal of uniformity in judicial decision-making. Some judges use COMPAS as a way to be informed while others use it as evidence. "Algorithmic risk assessment tools in the criminal justice context are models intended to predict a person's likelihood of committing a future crime. In practice, however, these tools may be used in 'off-label' ways, such as to help determine the length of a sentence."[8] When judges using COMPAS might treat defendants with high-risk scores more harshly, even when that score is not a perfect predictor of recidivism. Also, judges' discretionary use (some used, some didn't) led to deployment bias.

9. Self-Selection Bias

It occurs when the individuals or data points that are included in the training data are not representative of the whole population because of voluntary participation or some systematic exclusion. For example, this occurs when one trains a medical diagnosis model using data from patients who voluntarily visited a hospital. This excludes people who are sick but never go to the doctor, leading to a biased understanding of the disease. Another example is when a survey model trained on people who chose to respond to an online poll. Non-respondents may have different opinions, race, socio-economic status, etc. This means that the training data is biased, not be fully representative, due to who chooses to participate or is able to participate. In contrast to deployment bias, self-selection bias is introduced **during training, not use**.

This bias emerges through voluntary participation for training (e.g., defendants who voluntarily fill surveys needed for training); behavioural bias (collecting data requires a behavior that disable individuals cannot do) or desirability bias (users give the information that benefits them).

In COMPAS, "Defendants may seek to provide answers that would decrease the appearance of being considered an ongoing threat" when they complete the questionnaire.[9] In this scenario, the defendant answers the survey question in a way that the defendant be viewed as less of a threat to the public. This is an example of where individuals share the data that they think is needed, that is better for them, or that is expected, and exclude other information.

[8] Suresh and Guttag, "A Framework for Understanding Sources."

[9] Yasser Rahrovani, "ML Bias: AI in the Courtroom," January 21, 2025, in *CaseCast*, podcast, MP3 audio, 42:56, https://www.buzzsprout.com/2293124/episodes/16470939. See Supplemental Materials.

Mitigation

This case does not focus on in-depth bias mitigation strategies. However, instructors can give a sense of what can be done at a high level:

- Preprocessing: Addressing biases in the data before training by techniques such as reweighting, resampling data, or using multiple complementary measures.
- In-processing: Modifying the learning algorithm itself to incorporate various goals, such as fairness considerations, in addition to predictive accuracy.
- Post-processing: Adjusting the output of a trained model to reduce disparities.

TEACHING PLAN

The case is suitable for two 80-minute classes based on the COMPAS podcast to articulate the ML development process and the three categories and nine types of bias. The following is a suggested time allocation:

Discussion topic	**Time (minutes)**
Class 1	
Introduction	5
Exercise 1: Who would you side with? ProPublica or Northpointe?	10
Exercise 2: Teamwork: Bias identification	45
Exercise 3: The machine learning development cycle	20
Class 2	
Exercise 4: Mapping biases in machine learning development	15
Exercise 5: What is meant by false positives and false negatives in ProPublica's argument?	40
Exercise 6: Bias mitigation responsibilities and strategies	10
Debrief	15
Total	**160**

ANALYSIS

Introduction

Instructors can start the class by emphasizing that ML bias has become one of the most critical and widely discussed topics in technology and society today. They can highlight how biases in ML systems impact real-world decisions in fields such as policing, health care, and hiring with consequences that are not only technical but deeply ethical and societal. Setting this tone will help students to appreciate the importance of understanding ML bias as they prepare to navigate these challenges in their professional lives.

Next, instructors should underscore the need for students to understand what happens "under the hood" in ML systems. Without a solid grasp of how these systems work, biases can go unnoticed, leading to decisions that perpetuate systemic inequalities. To illustrate the real-world implications, instructors can talk about or show any of two short videos.

- The first introduces the use of policing technology, showcasing both its potential for public safety and its risks when misapplied.[10]
- The second tells the story of a facial recognition system that wrongly identified an innocent Black citizen, devastating his life and highlighting the human cost of algorithmic bias.[11]

Last, instructors can explain that the session will focus on two key areas: 1) understanding how ML systems are developed; and 2) identifying the sources of bias that can arise in the process. They should emphasize that these insights will be valuable across any AI initiative, whether students work on applications in hospitals, mobile apps, or other industries. By framing the session in this way, instructors set the stage for an engaging and practical discussion that connects technical concepts to real-world impact.

While COMPAS is commonly referred to as an AI system, some critics argue this label is a misnomer. The underlying model may be as simple as a logistic regression or even a rules-based scoring system. If AI is defined by technical sophistication—such as deep learning, adaptive feedback, or neural networks—then COMPAS arguably falls short. However, if AI is defined by function—a system that automates or mimics complex human decision-making—then COMPAS clearly qualifies. While this definitional ambiguity offers a valuable teaching moment, we suggest leaving this topic for another class, in which we focus on “What is (not) AI.”

Whether or not COMPAS is "really AI" matters less for our analysis and discussion of ML bias. The machine learning development cycle—data, preparation, modeling, evaluation, deployment—is pretty the same. COMPAS development and deployment in that real world can create harms similar to many AI systems: bias, opacity, scale, and the illusion of objectivity. Even if COMPAS represents a case of so-called “snake oil” AI—a system that appears intelligent but is overpromises and does not deliver what it claims—it remains highly relevant. It helps us understand the stakes, failures, and responsibilities that accompany the deployment of predictive AI in any organizational context.

1. Who would you side with? ProPublica or Northpointe?

Instructors can open the discussion by asking this question. The purpose of this is not to reach a particular conclusion on this complex topic but to open up different perspectives. Students are likely to raise the following:

Reasons to side with ProPublica include

- disparate racial impact: higher false positives for Black defendants;
- real-world consequences: harsher treatment due to risk misclassification;
- critique of fairness definitions: Northpointe’s metrics do not address disparate impact;
- transparency issues: the algorithm is a black box (mystery and difficult to assess); and
- perpetuating historical bias: the model reflects existing societal disparities.

Reasons to side with Northpointe include

- the model being an objective and unbiased tool to reduce human bias; and
- predictive parity (or accuracy equity): there is equal recidivism likelihood for same risk scores across races.

[10] CNBC, “How AI Could Reinforce Biases in the Criminal Justice System,” March 18, 2019, YouTube video, 0:19, https://youtu.be/ZMsSc_utZ40?feature=shared. See Supplemental Materials.

[11] Sky News, “Racial Bias in AI: Man Wrongly Identified by Facial Recognition Technology,” November 10, 2023, YouTube video, 1:23, https://youtu.be/Cx284WjpEQY?feature=shared. See Supplemental Materials.

Without going into details, instructors can move on to the next exercise: a better understanding of the different biases that may arise in ML development cycle is necessary first.

2. Teamwork: Bias Identification.

Instructors can divide the class into small groups (3-4 students per group). The goal of this activity is to identify the different types of biases students spotted in the podcast.

Each group is assigned one of the three bias categories, making sure that at least 2–3 groups are working on each category:

- biases in the data;
- biases in the algorithm; and
- biases in users (behaviours and population).

Allow 15 minutes for group discussion and identification of specific instances of biases within their assigned category. Then, the instructor opens the floor for another 30 minutes (more if needed) for the groups to articulate these types of biases. Instructors should ensure to discuss each bias and to caption them properly. This requires going back and forth to find examples, articulate criteria, and contrast with other biases. Instructors should write each bias on the board under the relevant categories in three columns: data, algorithm, and users. Based on my experience of teaching this class, the class always missed a few biases and sometime misplaces their category (user, data, or algorithm). Properly labeling, defining, and classifying these biases will make the discussion engaging and filled with learning.

Instructors can finish this exercise by emphasizing that there may be other subtypes or different ways of categorization. However, this list is sufficient to provide a sense of the main types of bias.

3. The machine learning development cycle.

After listing the nine biases, instructors should ask students to identify where and when they emerged in order to map them in the ML development cycle. Based on what students heard in the podcast, the instructor can ask “how did the development of COMPAS start?” Instructors can use students’ comments to emergently build a figure on the board (see Exhibit TN-1).

4. Mapping biases in machine learning development.

Instructors can then ask students to map the biases that they articulated in the previous exercise. Instructors can begin by mapping students’ identified biases on the board using different colours: this where each of the biases emerge on the ML development cycle. As shown in Exhibit TN-1, the nine biases are mapped on the ML development cycles in three colors: representation, measurement, and aggregation biases in orange; user-interaction, learning, and evaluation bias in light blue; deployment, historical, and self-selection in black. As a result, student can see the locus of each bias on the cycle.

Following this, instructors can ask students to identify how they see the “relationship” between the three bias categories (data, user, and algorithm) and sketch this out, emphasizing the importance of mitigating this relationship in any AI or ML project (see Exhibit TN-2). This circular relationship between categories of biases can be seen in the ML development cycle too.

5. What is meant by false positives and false negatives in ProPublica’s argument?

With a better view of the big picture of ML development, types of biases, and the locus of where they are created, it is time to go back to the initial debate about COMPAS’s bias and evaluate the debate. Before discussing the fairness types and their tension, instructors should ensure students understand what false positives and false negatives are:

- False negative is the prediction that an individual will not reoffend when they do.
- False positive is the prediction that an individual will reoffend when they do not.

Instructors should approach the ProPublica versus Northpointe arguments in two ways, by considering:

1. Those who did or did not reoffend and then calculates what percentage was incorrectly or correctly labelled, respectively, as high-risk (which is what ProPublica did).
2. Those who were labelled as either low-risk or high-risk by the algorithm and then compare the result against the rate at which they did or did not, respectively, reoffend (which is what Northpoint did).

After understanding what the concepts are, instructors show on the board (if do not have time) or can ask students to go deeper into understanding the core argument of the two sides of the debate. If students are statistics-savvy, instructors can assign half the groups to calculate the numbers for ProPublica’s arguments and the other half to Northpointe’s.

Instructors can then move on to help students better understand the crux of the argument on both sides.

ProPublica’s Argument: Fairness in the Outcome

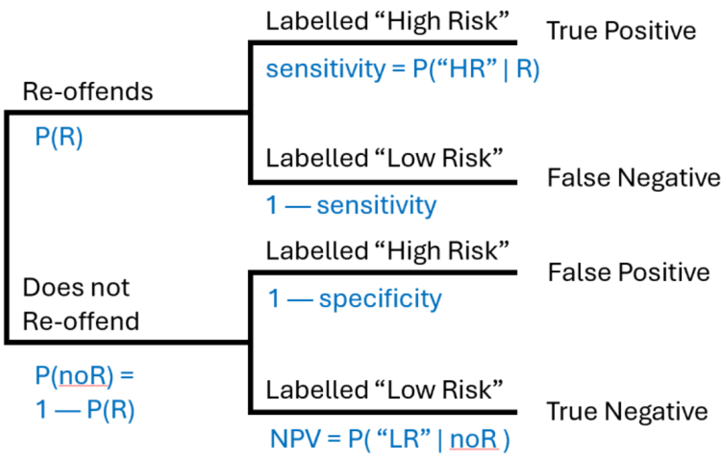
Instructors can summarize the argument for ProPublica by showing the following statistical figures and tables. We have prepared a slide deck that has the tables and Excel sheet that shows the calculations.

ProPublica identified that the false negative rate (FNR) and false positive rate (FPR) are differentiated by race. Using a probability tree, looking at those who reoffended, errors in mistakenly flagging someone as low-risk were almost double for white defendants (0.477) compared to Black defendants (0.280). This led to significantly lower penalties for reoffending white defendants. Alternatively, looking at those who did not reoffend, errors in mistakenly flagging someone as high-risk were almost double for Black defendants (0.449) compared to white defendants (0.235). This led to significantly harsher penalties for non-reoffending Black defendants.

Table 1: ProPublica’s probability of false negative and false positive rates

Label	Probability	Black defendants	White defendants
Positivity rate	P(“HR”)	0.587	0.347
True positive rate (TPR)	P(“HR”|R)	0.720	0.523
True negative rate (TNR)	P(“LR”|noR)	0.551	0.765
False negative rate (FNR)	1-P(“HR”|R)	0.280	0.477
False positive rate (FPR)	1-P(“LR”|noR)	0.449	0.235

Note: Note: HR = High Recidivism label by COMPAS; R = Recidivism happened in reality; LR = Low Recidivism label by COMPAS; noR = No recidivism happened in reality.
P("HR"): Probability of one being labeled as High Recidivism (HR)
P("HR"|R): Probability of one being labeled as High Recidivism by COMPAS, and they did recidivism (correct labeling)
P("LR"): Probability of one being labeled as Low Recidivism (LR)
P("LR"|noR): Probability of one being labeled as Low Recidivism by COMPAS, and they did not do recidivism (correct labeling)



Note: HR = High Recidivism label by COMPAS; R = Recidivism happened in reality; LR = Low Recidivism label by COMPAS; noR = No recidivism happened in reality

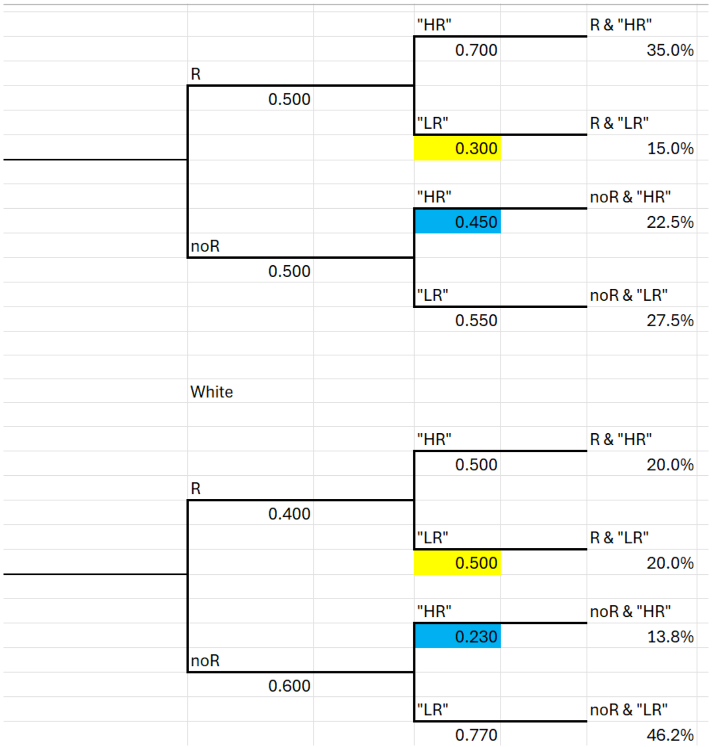
To calculate and better understand these numbers, instructors can show Table 2 with three probable outcomes and then calculate the probability of the four scenarios, as the resulting numbers illustrated in the second probability tree. The table is include in he slide deck.

Table 2: Three probable outcomes in ProPublica's calculations

Scenario	Black defendants	White defendants	Probable outcome
P(R)	0.50	0.40	Probability defendant reoffends within two years.
P ("HR" | R)	0.70	0.50	Sensitivity. Also called the true positive rate. Probability that the test will correctly assign someone who will reoffend within two years as high-risk.
P ("LR" | noR)	0.55	0.77	Specificity. Also called the true negative rate. Probability that the test will correctly assign someone who will not reoffend within two years as low-risk.

Note: HR = High Recidivism label by COMPAS; R = Recidivism happened in reality; LR = Low Recidivism label by COMPAS; noR = No recidivism happened in reality

Putting the values from Table 2 into the second probability tree will result in values (highlighted in blue and yellow) that are close to those in Table 1.



Note: HR = High Recidivism label by COMPAS; R = Recidivism happened in reality; LR = Low Recidivism label by COMPAS; noR = No recidivism happened in reality

To recap, instructors can conclude with what fairness is in ProPublica's argument

Based on the above analysis, instructors can summarize what it means to be fair in terms of false negative or false positive rates. *It is fair when the false positive and false negative errors are close for both Black and white defendants*, i.e., fairness in the outcome.

- False positive rate balance: This condition mandates that the rate at which non-reoffenders are incorrectly labelled as high-risk should be the same for different racial groups. In other words, the proportion of false positives should be equal across races.
- False negative rate balance: This condition requires that the rate at which reoffenders are incorrectly labelled as low-risk should be equal across racial groups. This means that the proportion of false negatives should be the same for both Black and white defendants.

This is called error rate balance or fairness in the outcome.

Northpointe's Argument: Proponent of Fairness in the Procedure

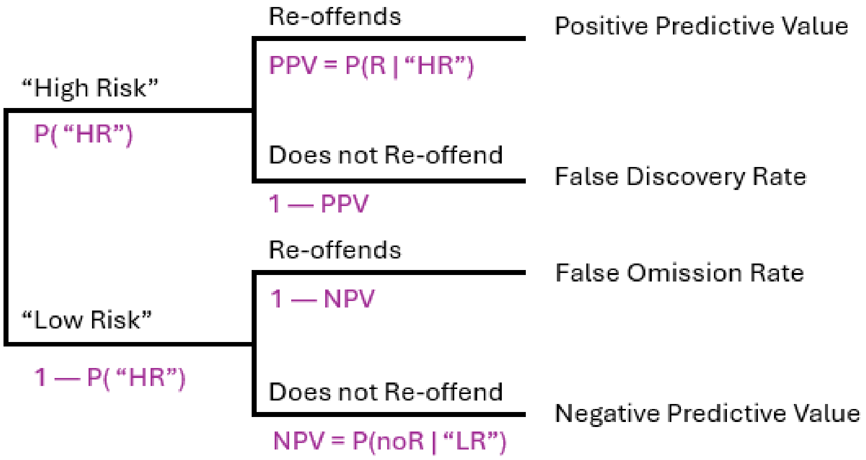
Instructors can summarize Northpointe's counter-argument by showing the following statistical figures and tables.

Northpointe counter-argued that the positive predictive value (PPV) and, therefore, the false discovery rate (FDR) are very similar by race. As evidenced in the following probability tree, Northpointe argues that looking at those the algorithm has labelled as high-risk and checking against whether they did reoffended or not shows that Northpointe's predictive value is almost the same across races (0.63 for Black and 0.59 for white defendants). Northpointe also argues that its labelling is accurate and equally predictive of reoffending across races. Therefore, its FDR is also almost identical (1-PPV): 0.37 for Black and 0.41 for white defendants.

Table 3: Northpointe's probability of reoffending

Label	Probability	Black defendants	White defendants
Base rate	P(R)	0.510	0.390
Positive predictive value (PPV)	P(R|"HR")	0.630	0.590
Negative predictive value (NPV)	P(noR|"LR")	0.650	0.710
False discovery rate (FDR)	1-P(R|"HR")	0.370	0.410
False omission rate (FOR)	1-P(noR|"LR")	0.350	0.290

Note: HR = High Recidivism label by COMPAS; R = Recidivism happened in reality; LR = Low Recidivism label by COMPAS; noR = No recidivism happened in reality
P("R"): Probability of one doing Recidivism, that is the base rate, reality in a society
P(R|"HR"): Probability of recidivism among those who were labelled by COMPAS as "High Recidivism" (correct prediction)
P(noR|"LR"): Probability of no recidivism among those who were labelled by COMPAS as "Low Recidivism" (correct prediction)



Note: HR = High Recidivism label by COMPAS; R = Recidivism happened in reality; LR = Low Recidivism label by COMPAS; noR = No recidivism happened in reality

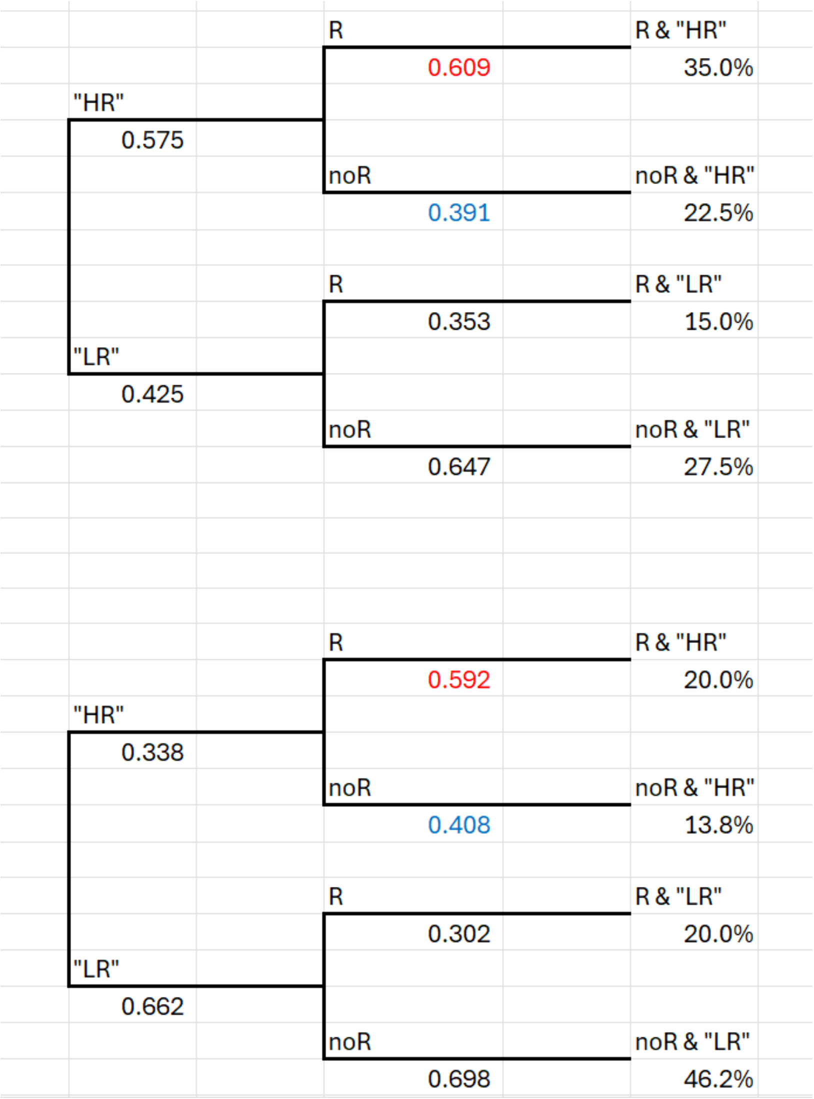
To calculate and better understand the numbers in Table 3, instructors can show Table 4 with three probabilities and then calculate the probability of the four scenarios—as illustrated in the fourth probability tree.

Table 4: Three probable outcomes in Northpointe's calculations

Scenario	Black defendants	White defendants	Probable outcome
P("HR")	0.575	0.338	Probability of the test indicating a high risk (e.g., >= 5).
P (R | "HR")	0.609	0.592	Positive predictive value (PPV). Probability that someone identified as high-risk will reoffend within two years.
P (noR | "LR")	0.647	0.698	Negative predictive value (NPV) Probability that someone identified as low-risk will not reoffend within two years.

Note: HR = High Recidivism label by COMPAS; R = Recidivism happened in reality; LR = Low Recidivism label by COMPAS; noR = No recidivism happened in reality.

Putting the values from Table 4 into the fourth probability tree will result in values (coloured blue and red) that are close to those in Table 3.



Note: HR = High Recidivism label by COMPAS; R = Recidivism happened in reality; LR = Low Recidivism label by COMPAS; noR = No recidivism happened in reality

To recap, instructors can start by asking: "What is fairness in Northpointe's argument?"

Based on the above analysis, instructors can summarize it as Northpointe arguing that COMPAS was equally predictive for both Black and white defendants and dismissed the racial disparities detected by ProPublica as irrelevant.

This condition requires that individuals classified as high-risk should have the same likelihood of reoffending regardless of their race—in other words, PPV and FDR are close across races—demonstrating predictive parity or fairness in the procedure.

Conflict Between Fairness Definitions

Students eventually appreciate that it is impossible to simultaneously satisfy all these fairness conditions because Black defendants have a higher overall recidivism rate compared to white defendants in the context of ML training data. Here are the main arguments:

- Differential base rates: The higher recidivism rate for Black defendants (51 per cent in Broward County compared to 39 per cent for white defendants) means that the base rate of reoffending is different between the two racial groups (i.e., there is historical bias in the data). This disparity creates a fundamental tension in achieving fairness.
- Fairness types: Predictive parity (procedure fairness) versus error rate balance (outcome fairness). To achieve predictive parity, the risk scores must reflect the higher base rate of reoffending among Black defendants. However, this often leads to higher false positive rates for Black defendants or higher false negative rates for white defendants. Conversely, balancing false positive and false negative rates across races would require adjusting the thresholds in a way that compromises predictive parity.
- Trade-offs in fairness: Given the different base rates (societal, historical bias), achieving one type of fairness often exacerbates disparities in another. For instance, equalizing false positive rates might result in predictive parity being skewed, as the higher recidivism rate among Black defendants necessitates different risk thresholds for accurate prediction.

6. Bias mitigation responsibilities and strategies.

Although the case is not about mitigation strategies, if time allows, instructors can raise the question with students around whose responsibility it is to de-bias and how they would propose to achieve that.

Overall, everyone is responsible for bias, as everyone has a part in creating it within the vicious circle. Northpointe needs to incorporate outcome fairness, and the US judicial system needs to be mindful of societal biases beyond Northpointe’s control and to have a long-term plan in place for mitigating them. In addition, the US judicial system cannot adopt and implement an AI system that is mystery (blackbox) in how it operates. There needs to be greater understanding and knowledge around data collection, variable measurement, model training, internal and external validation, and of the algorithm itself, to the extent that judges know what they are applying. This will affect training, retraining for deployment in new contexts, and maintenance. This improved transparency can be done with the help of independent research to ensure copyright protection.

Several practices can be adopted to reduce bias, although it cannot be eliminated. Students are likely to raise the following:

Data-Driven Mitigation

- Reweighting or resampling data. Adjusting the data set to ensure that minority groups are represented proportionally during the training phase. This could help reduce representation bias and make predictions more equitable across racial groups.
- Refining proxy variables. Replacing problematic proxies like “arrest” or “rearrest” with more direct measures of crime and recidivism (e.g., community supervision data, self-reported outcomes) would address measurement bias. Remember that there is no perfect measure.

Algorithmic Adjustments

- Fairness-constrained optimization. Incorporating fairness constraints into the algorithm, such as limiting the disparity between false positive rates across racial groups. This would ensure the algorithm accounted for error rate balance during training.
- Subgroup-specific models. Developing separate models for different demographic groups to handle aggregation bias. For instance, having tailored models for different communities might prevent the one-size-fits-all issue.
- Transparent feature selection. Clearly explaining the choice of input features and their contribution to predictions. This can reduce hidden biases in the algorithm and address criticisms of using opaque methods.

Post-processing

- Score adjustment. Modifying output scores to ensure that decisions (e.g., parole eligibility) improve fairness objectives, such as narrowing the gap on error rates across racial groups. Techniques such as threshold adjustment can help to mitigate disparate impact.
- Calibration across subgroups. Recalibrating risk scores to ensure they align with observed recidivism rates within each racial group, without sacrificing predictive parity. For instance, adjusting thresholds for classifying high risk differently for different demographic groups.

Evaluation and Validation

- External validation across diverse groups. Testing the model on external data sets representing diverse populations to ensure robustness and fairness in predictions across different racial groups.
- Incorporating multi-metric evaluations. Going beyond overall accuracy and including fairness metrics (e.g., demographic parity, error rate balance) in performance evaluations.

Improving User Interaction

- Redesigning interfaces. Presenting risk scores in ways that minimize misinterpretation, such as including contextual explanations or confidence intervals. For example, instead of presenting scores as binary outcomes, using a probabilistic or gradient scale to reduce confirmation or automation bias.
- Training stakeholders. Providing judges and other users with guidance on how to interpret and use COMPAS scores appropriately, ensuring that the tool informs rather than dictates decisions.

Data Transparency

- Demystifying the algorithm. Making the COMPAS algorithm more transparent by publishing details about how it operates and what features it prioritizes, at least for the organizations that use it. Transparency helps to build trust and allows external audits to identify potential biases.
- Feedback loops. Collecting feedback from users (e.g., judges, probation officers) to identify cases where the tool’s recommendations deviate significantly from observed outcomes, and using this feedback to refine the model iteratively.

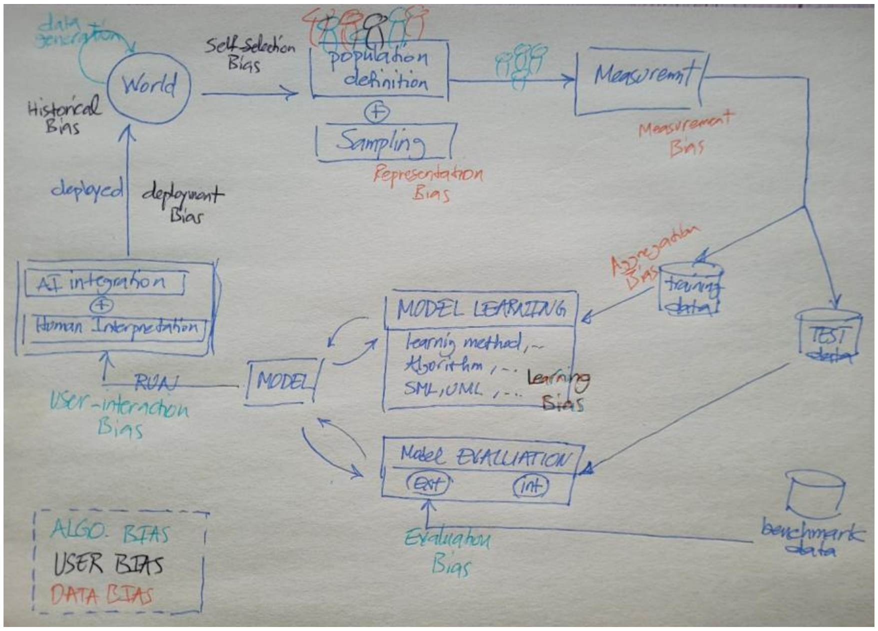
These strategies highlight that ProPublica’s concerns require foundational changes to data and fairness considerations, while Northpointe must improve transparency, user interaction, and its consideration of error rate disparities. Balancing these perspectives is essential for designing fair and robust ML systems like COMPAS.

DEBRIEF FOR ARTIFICIAL INTELLIGENCE INITIATIVES

Instructors can now bring the conversation back to a whole-class discussion to articulate the implications for AI and ML project development and implementation. The following points should come up:

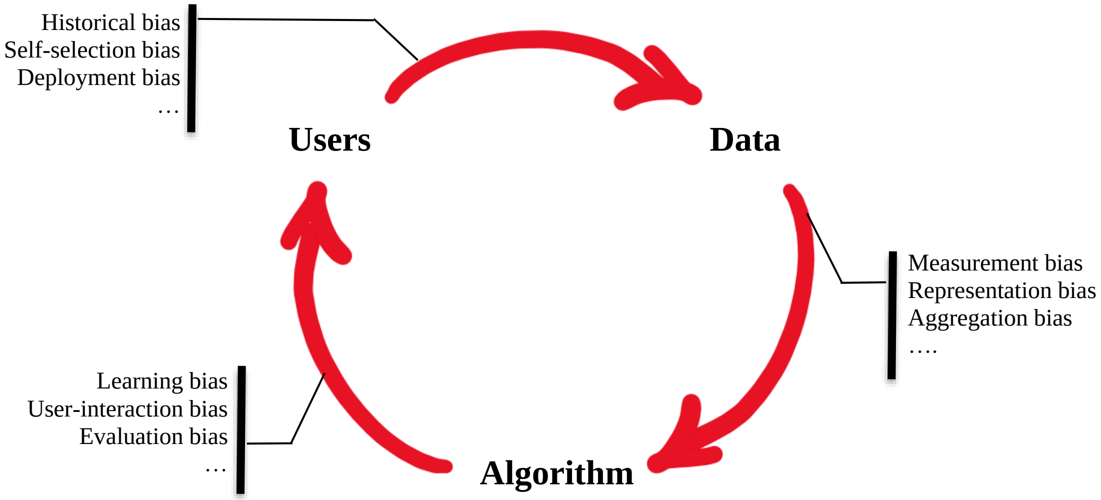
- The importance of understanding the ML development cycle, different types of biases, and the locus of bias creation.
- An acknowledgement of the circular relationship between biases—from data to algorithm to user, and back to data.
- The complex interplay between different definitions of fairness in AI initiatives. What is the definition of fairness in a project, and why it needs to be mindfully addressed and pre-planned.
- The significant impact of the choices in ML development (e.g., in measurement, training data, internal and external validation, algorithms). There is no perfect measure. Any measure can create some biases that are scaled with a widely deployed AI system.
- The need for a nuanced approach to fairness that recognizes the inherent trade-offs, entertains various definitions and impacts, and prioritizes ethical and transparent decision-making in the criminal justice system. The impossibility of satisfying all fairness criteria simultaneously necessitates careful consideration of which fairness measures to prioritize. This involves ethical judgments about the relative importance of different types of fairness and the potential impacts on different racial groups.
- As not all trade-offs can be resolved, transparency is necessary in the choices made before deploying AI tools. Transparency and accountability in ML development and the implementation process, particularly around the design and application of risk assessment tools. Policy-makers and practitioners need to be aware of the trade-offs and make informed decisions about the use of these tools within the justice system.
- AI systems cannot be set up and forgotten about. There is an ongoing refinement loop as biases scale and interact. Systems need to be constantly monitored, retrained, and de-biased, given the bias circularity and scale of system deployment. AI systems need maintenance, as data feeds back in and can magnify some biases and, as such, need to be monitored across a diverse variety of fairness measures.
- It is important to identify who the AI project stakeholders and beneficiaries are (e.g., data scientists, ethicists, judges, defendants, the public, government, non-governmental organizations). They are collectively responsible for AI's pitfalls and promises.
- Retraining is necessary over time and especially when a product is deployed into a new context.

EXHIBIT TN-1: THE MACHINE LEARNING DEVELOPMENT CYCLE



Source: Created by the case authors based on Harini Suresh and John Guttag, “A Framework for Understanding Sources of Harm Throughout the Machine Learning Life Cycle,” *Equity and Access in Algorithms, Mechanisms, and Optimization*, EEAMO ’21: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 1–9. https://doi.org/10.1145/3465416.3483305.

EXHIBIT TN-2: THE MACHINE BIAS VICIOUS CIRCLE



Source: Created by the case authors based on case information.