The COMPAS case: an educational journey for explaining fairness in AI-based applications

Antonio Rodà[1,*]

[1]Department of Information Engineering, University of Padova, via Gradenigo 6a, 35131, Padova, Italy

Abstract

This article explores the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) case as an invaluable educational resource for elucidating the complexities of fairness assessments in real-world AI applications. The controversial history of this high-profile legal and media case is briefly reviewed, highlighting the multifaceted nature of algorithmic bias in criminal justice. By examining the various analytical approaches proposed to address the COMPAS algorithm's fairness, this paper underscores the limitations and challenges inherent in each methodology. The case study serves as a compelling illustration of the intricate balance between statistical accuracy, social justice, and ethical considerations in AI-driven decision-making systems. Through this analysis, the article aims to provide students and practitioners with a nuanced understanding of the practical difficulties in achieving and measuring fairness in AI applications.

Keywords

LaTeX class, paper template, paper formatting, CEUR-WS

1. Introduction

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) case has been extensively documented and analyzed from various perspectives, generating a substantial body of literature (see e.g., [1, 2, 3, 4, 5]). It has become a focal point in discussions about algorithmic fairness, transparency in artificial intelligence (AI) systems, and the ethical implications of using predictive tools in the criminal justice system.

In this article, I aim to look at this case from another point of view: the pedagogical value of the COMPAS case study, as an invaluable educational resource for elucidating the complexities of fairness assessments in real-world AI applications. This work stems from the experience of teaching a course titled Gender Knowledge and Ethics in Artificial Intelligence [6], which has been active for a few years at the University of Padua. I will present materials and an analytical framework that considers the viewpoints of the key stakeholders (ProPublica, Northpointe, and the Wisconsin Supreme Court), critically reviewing their conclusions. This approach is designed for students of a bachelor degree in computer science/engineering and aims to systematically deconstruct hasty judgments, encouraging students to exercise critical thinking skills. The ultimate goal is to foster in students an awareness that only a perspective that embraces the complexity of real-world problems can lead to robust analyses that are less prone to error. By engaging with the multifaceted nature of the COMPAS controversy, students will develop a nuanced understanding of the challenges in algorithmic fairness and the importance of considering multiple dimensions when evaluating AI systems in sensitive domains such as criminal justice.

The remainder of this paper is structured as follows. After providing a brief summary of the COMPAS controversy (Section 2), reviewing the key events and positions taken by the major stakeholders - ProPublica, Northpointe, and the Wisconsin Supreme Court, I will outlines an educational journey (Section 3) through the COMPAS case, presenting an analytical framework that systematically deconstructs the arguments and evidence put forth by each party.

*Corresponding author.
antonio.roda@unipd.it (A. Rodà)
https://dei.unipd.it/~roda (A. Rodà)
0000-0001-9921-0590 (A. Rodà)

2. A brief summary of the COMPAS controversy

COMPAS is a risk assessment tool developed by Northpointe (now Equivant) to assist courts in making decisions about pretrial release, sentencing, and parole [7]. COMPAS uses an algorithm to predict an offender's likelihood of recidivism based on various factors, including criminal history, substance abuse, and social-familial risk [8].

In May 2016, the investigative journalism organization ProPublica published a groundbreaking article titled "Machine Bias" [9]. This exposé alleged that the COMPAS algorithm was biased against African American defendants, claiming that the tool was almost twice as likely to falsely label black defendants as future criminals compared to white defendants.

In response to ProPublica's allegations, Northpointe issued a technical report challenging ProPublica's methodology and conclusions [10]. The company argued that their tool satisfied statistical fairness criteria, specifically emphasizing equal positive and negative predictive values across racial groups.

The controversy surrounding COMPAS reached a critical point in the case of Loomis v. Wisconsin in 2016. Eric Loomis, sentenced partly based on his COMPAS score, challenged the use of the algorithm in sentencing, arguing that a) it violated due process and b) relied on gender-based assessments. In particular, the appellant argued that he had been subjected to discrimination on the basis of his gender, specifically because he is male. In July 2016, the Wisconsin Supreme Court ruled on the Loomis case, upholding the use of COMPAS in sentencing while also highlighting concerns about its proprietary nature and potential biases[11]. The court mandated that judges be informed of the tool's limitations, including its potential for gender and racial bias, when considering COMPAS scores in sentencing decisions.

3. An educational journey

3.1. Step 1: From Anecdotal Evidence to Statistical Analysis

The educational pathway begins by examining ProPublica's exposé on the COMPAS system. Students are initially presented with a powerful visual juxtaposition, directly taken from the ProPublica's article[1]: a photograph of an African American woman who received a high-risk score from COMPAS, alongside an image of a Caucasian man who was assigned a low-risk score, although they were both arrested for a petty theft. This striking contrast serves as a compelling entry point into the discussion of algorithmic bias. When asked to interpret these images, students often swiftly conclude that they are witnessing a clear-cut case of racial discrimination. This immediate reaction provides an excellent opportunity to caution them against drawing hasty conclusions based on anecdotal evidences, no matter how emotionally impactful it may be.

To challenge this initial perception, it is crucial to present counterexamples. These might include cases, extracted from the same database used by ProPublica, of recidivate African Americans who received low-risk scores, and Caucasians not recidivate who were assigned high-risk scores, as in Figure 3.1. This balanced approach helps students understand that individual cases, while powerful, can be misleading when assessing the fairness of complex systems like COMPAS. Through this exercise, students begin to grasp that determining unfairness in algorithmic decision-making systems requires more than anecdotal evidence. It necessitates a rigorous statistical approach that can account for the complexity and scale of the data involved.

3.2. Step 2: From distribution analysis to societal context

The educational journey continues by examining the distribution of risk scores assigned by COMPAS, ranging from 1 (low risk) to 10 (high risk), segregated by race as presented in ProPublica's article [2]. The

[1] Look at the Vernon Prater and Brisha Borden pictures in https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed on February, 15th 2025.

[2] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed on February, 15th 2025.

Figure 1: Names and data are extracted from the same dataset used by ProPublica. Robert was arrested for petty theft and received a high-risk score, although no subsequent crimes were recorded in the dataset. On the contrary, Nathan received a low-risk score, but he was arrested again some months later. The pictures were generated by OpenArt.AI, giving the following prompts: "Create a photo of a 38-year-old Caucasian male in the style of police mugshots" and "Create a photo of a 55-year-old African American male in the style of police mugshots".

visual representation reveals stark differences between Caucasian and African American populations. For African Americans, the distribution shows a gradual decline, with approximately 370 individuals receiving a score of 1 and about 220 receiving a score of 10. In contrast, the distribution for Caucasians exhibits a steeper decline, starting with around 600 cases scoring 1 and dropping to merely 50 cases scoring 10. The disparity between these distributions is conspicuous, usually leading students to conclude that the system appears discriminatory towards African Americans, also from a statistical point fo view.

These data provide an opportunity to broaden the students' perspective. It's crucial to emphasize that what was shown aren't just numbers, but represent real people living within specific social, economic, and cultural contexts. In this case, the data pertains to individuals in a world still grappling with significant racial disparities. To contextualize these findings, students are presented with statistics highlighting racial disparities in the United States, such as those reported by the Economic Policy Institute (EPI), an independent think tank that researches the impact of economic trends and policies on working people in the United States. According to the EPI's study on racial disparities [3]

- the median household income in 2023 was $56,490 for black people and $89,050 for white people
- in 2023, 12.6% of black people has less than a high school education, whereas this percentage drops to 6.1% for white people
- black women are over twice as likely to die from a pregnancy-related cause (49.5 deaths per 100,000 births) as white women (19.0 deaths per 100,000 births)
- over 1,000 out of every 100,000 U.S. residents who are Black or were imprisoned in 2023, more than four times the proportion related to white U.S. residents (229 out of 100,000)

This broader view encourages students to consider how societal inequities might be reflected in the risk assessment scores. At this point, an alternative hypothesis can be introduced: the distribution differences in the risk scores between Caucasians and African Americans might be due to an actual

[3] https://www.epi.org/publication/disparities-chartbook/

Table 1
Confusion matrix from the public dataset on COMPAS

Black defendants		
	Low	High
Survived	990	805
Recidivated	532	1369

White defendants		
	Low	High
Survived	1139	349
Recidivated	461	505

higher risk of the latter, suggesting that COMPAS could be accurately performing its intended function. To test this hypothesis, the analysis must move beyond descriptive statistics. Students are guided to consider the system's performance in terms of the errors it commits. This transition sets the stage for a more nuanced examination of fairness metrics, emphasizing the importance of evaluating algorithmic systems not just on their outputs, but on their accuracy.

3.3. Step 3: Unveiling the Complexities of Fairness Metrics

At this third stage of our educational journey, it becomes crucial to analyze data that reflect COMPAS's predictive capabilities. Table 1 presents the confusion matrices separated by race for African Americans and Caucasians, as reported by ProPublica [4]. These data were calculated from a public database [5], containing 7214 entries with criminal history, jail and prison time, demographics and COMPAS risk scores for defendants from Broward County in the period 2013-2014. The author has independently verified the accuracy of ProPublica's reported values. To limit the complexity of the discussion, which is of little use for didactic purposes at this stage, I will only use the fairness metrics considered by ProPublica. An exhaustive treatment of fairness metrics can be proposed to students at a later stage or as personal study, after properly motivating them on the usefulness of considering multiple metrics.

In this context, 'recidivated' refers to individuals who were re-arrested within two years of receiving their risk score, while 'survived' indicates those who avoided arrest during the same period. Let's first examine the 'survived' row, representing individuals who "behaved well" and for whom COMPAS was expected to predict a low risk. The False Positive Rate (i.e. people "unjustly" assigned a high risk) shows a significant disparity:

- For African Americans: 44.85% of survived individuals (calculated as $805/(805+990)\times 100$)
- For Caucasians: 23.45% of survived individuals (calculated as $349/(349+1139)\times 100$)

Now, let's turn our attention to the 'recidivated' row. The False Negative Rate (i.e. people "unjustly" assigned a low risk) also reveals a notable difference:

- For African Americans: 27.99% of recidivated individuals (calculated as $805/(805+990)\times 100$)
- For Caucasians: 47.72% of recidivated individuals (calculated as $349/(349+1139)\times 100$)

These findings led ProPublica to conclude:

> "Black defendants who do not recidivate were nearly twice as likely to be classified by COMPAS as higher risk compared to their white counterparts (45 percent vs. 23 percent). White defendants who re-offended within the next two years were mistakenly labeled low risk almost twice as often as black re-offenders (48 percent vs. 28 percent)."

[4] https://www.ProPublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Accessed on February, 15th 2025
[5] https://github.com/propublica/compas-analysis

At this juncture, the data on False Positive Rate (FPR) and False Negative Rate (FNR), as summarized in the aforementioned quote, seem to leave little doubt about COMPAS's discriminatory behavior towards African Americans. However, it is crucial to guide students towards the understanding that fairness metrics can be deceptive if not fully comprehended in context.

Specifically, the disparate FPR values between Caucasians and African Americans do not necessarily imply discrimination. To demonstrate this, we present students with a counter-example, by using a simplified risk prediction system, which I'll call MYCOMPAS.

Consider two populations, labeled 'Green' and 'Purple', each comprising 100 individuals. We will examine the following two scenarios.

Scenario 1: Ideal World (No Bias).

In this scenario, we assume an equal proportion of survived and recidivated people in both populations, specifically 80 survived and 20 recidivated. MYCOMPAS was designed to behave "perfectly" equitably, according demographic parity, and assigns a high risk score to 20 individuals, making 5 errors, for both Green and Purple populations.

The confusion matrices for this scenario are presented in Table 2. For both populations, we observe

$$\begin{aligned} FPR_{Green} &= 5/(75+5) = 6.25\% \\ FPR_{Purple} &= 5/(75+5) = 6.25\% \end{aligned} \tag{1}$$

as expected given that MYCOMPAS is fair by design and there are no disparities between populations.

Scenario 2: Unequal World.

In this scenario, we introduce a disparity in the proportion of recidivated people. In particular, we assume that the Green population has a greater proportion of recidivated people (60 survived, 40 recidivated) in comparison to the Purple one (80 survived and 20 recidivated), i.e. Green people are more prone to reoffend. MYCOMPAS maintains its equitable behavior, assigning a high risk score to 20 individuals and making 5 errors for both populations. The resulting confusion matrices are shown in Table 3. From these matrices, we derive:

$$\begin{aligned} FPR_{Green} &= 5/(55+5) = 8.33\% \\ FPR_{Purple} &= 5/(75+5) = 6.25\% \end{aligned} \tag{2}$$

In this second scenario, therefore, we obtained that FPR is higher for Green population than Purple one. Given that MYCOMPAS is fair by design, it can be concluded that a disparity in FPR does not imply a discriminatory behavior of the predictive algorithm. Therefore, the difference in FPR arises not from discriminatory behavior of the algorithm, but from the underlying differences in the recidivism rates between the two populations. In other words, FPR depends on the ratio between False Positives and the number of people who do not commit further crimes (survived). When the latter decreases, FPR increases.

Going back to COMPAS now, the higher FPR value in African Americans does not necessarily imply a higher number of False Positives (which would suggest discrimination against African Americans), but rather a lower number of survived individuals.

Using the FPR metric to argue racial discrimination of COMPAS, as done in the ProPublica document, is therefore fallacious. The picture becomes even more complex when considering the Positive Predicted Value metric, also correctly reported by ProPublica, although not carefully examined. This metric is equal to the fraction of recidivated people among those who received a high score. A low value implies that COMPAS was unnecessarily severe, harming many people with a high risk score who then behaved well. From Table 1, we obtain:

$$\begin{aligned} PPV_{Black} &= 1369/(805+1369) = 0.63 \\ PPV_{White} &= 505/(349+505) = 0.59 \end{aligned} \tag{3}$$

Table 2
First scenario: ideal world (without bias)

Green defendants		
	Low	High
Survived	75	5
Recidivated	5	15

Purple defendants		
	Low	High
Survived	75	5
Recidivated	5	15

Table 3
Second scenario: more realistic world (with bias)

Green defendants		
	Low	High
Survived	55	5
Recidivated	25	15

Purple defendants		
	Low	High
Survived	75	5
Recidivated	5	15

revealing that PPV is larger for African Americans compared to Caucasians. In other words, this metric supports the conclusion that COMPAS was slightly more severe with Caucasians than African Americans.A conclusion that is the almost opposite to that of ProPublica.

The counter-example based on the MYCOMPAS model serves to caution students against drawing hasty conclusions based on a single fairness metric. It underscores the importance of considering multiple perspectives and metrics when evaluating the fairness of an AI system.

3.4. Step 4: From racial bias to gender discrimination

After deconstructing ProPublica's allegations and questioning their methodology and conclusions, we turn our attention to the document produced by Northpointe (now Equivant) [10]. This document, using different arguments from those presented in the previous sections, concludes that COMPAS does not produce discriminatory outputs against African Americans. This document can be presented to students, noting that its analysis is more comprehensive than ProPublica's and more rigorous from a methodological standpoint. At this juncture, many students may concur with the notion that COMPAS is, on balance, fair.

However, it is crucial to challenge this perception as well, introducing the concept of intersectionality. Intersectionality, as defined by Kimberlé Crenshaw (1989), refers to the interconnected nature of social categorizations such as race, class, and gender, regarded as creating overlapping and interdependent systems of discrimination or disadvantage.

To avoid overcomplicating the analysis, let's consider only the gender dimension. Table Z, which shows confusion matrices derived from the same public database used by ProPublica and Northpointe, but this time segmented by gender. Recalculating the Positive Predictive Value (PPV) metric for male and female groups yields the following:

Table 4
Second scenario: more realistic world (with bias)

Male defendants

	Low	High
Survived	1813	788
Recidivated	909	1487

Female defendants

	Low	High
Survived	532	230
Recidivated	167	246

$$\begin{aligned} PPV_{Male} &= 1487/(788+1487) = 0.65 \\ PPV_{Female} &= 246/(230+246) = 0.52 \end{aligned} \tag{4}$$

The resulting values of PPV, that are lower for females than males, indicate that COMPAS was more severe with women than with men, that is, of all the women who received a high risk score, only 52% went on to commit further offenses. Therefore, while ProPublica may not conclusively demonstrate COMPAS's discrimination against African Americans, it is also not possible conclude that COMPAS is simply fair, as asserted by NorthPointe: some metrics, PPV in particular, indicate indeed a potential bias against women. It is surprising that public debate has focused solely on racial discrimination, while discrimination against women has been overlooked. This oversight is even more striking considering that one of the contentions in the Loomis v. Wisconsin case was alleged discrimination against the appellant as a male. Paradoxically, our data analysis shows the opposite - females appear to receive "harsher" treatment from the COMPAS system.

These findings underscore the necessity for considering different (possible all) social categorizations in fairness analysis. Possibly adopting an intersectional approach that considers multiple dimensions of identity simultaneously. The intersectional perspective is crucial for a comprehensive understanding of algorithmic fairness, particularly in complex systems like COMPAS that have far-reaching implications in the criminal justice system.

While the debate surrounding COMPAS has primarily centered on racial bias, our analysis reveals a more nuanced picture of potential discrimination.

3.5. Step 5: From fairness metrics to human-AI interaction

After critically discussing both ProPublica's document, which aimed to demonstrate racist behavior in COMPAS, and Northpointe's assertion of non-discrimination, further useful discussions with students can arise by analyzing the Wisconsin Supreme Court's ruling in the Loomis case. The court held that the use of the COMPAS recidivism risk calculation software by an ordinary court in criminal proceedings does not violate the defendant's right to due process, thus siding with Northpointe. However, it is interesting to cite the passage where the judges are keen

> "to clarify that while our holding today permits a sentencing court to *consider* COMPAS, we do not conclude that a sentencing court may *rely* on COMPAS for the sentence it imposes."

The judges consider this distinction so important that they further specify:

> "Contrary to the manner in which the majority opinion sometimes employs 'consider' and 'rely', they are not interchangeable. 'Rely' is defined as 'to be dependent' or 'to place full confidence'. Therefore, to permit circuit courts to rely on COMPAS is to permit circuit courts to depend on COMPAS in imposing sentence. On the other hand, 'consider' is defined as 'to observe' or 'to contemplate' or 'to weigh'."

These sentences allow us to introduce one of the pillars for a responsible use of AI-based systems: the principle of *meaningful human control*. This final stage of the educational journey proposed in this article aims at explaining what this term means, highlighting possible different levels of control (a posteriori evaluation, human intervention in case of anomalous behavior, continuous control, etc.) and the conditions (first of all explainability) under which control can be considered meaningful.

Moreover, the distinction between 'consider' and 'rely' is related to human factors (what happens with a human in the loop?) and the main biases that characterize human-machine collaboration, foremost among them being automation bias [12, 13]. Automation bias can be defined as the propensity for humans to favor suggestions from automated decision-making systems over contradictory information made without automation, even when the latter is correct. Can we rule out that a judge, dealing with an overload of decisions to make and sentences to issue, decides to trust (rely) the machine without exercising an effective critical control? What kind of training and awareness of the machine's limitations should a judge have to reduce this risk? What features should the interaction between judge and machine have to reach this aim?

It is important to note that the distinction between 'consider' and 'rely' can be extended to numerous other applications. For instance, consider the case of generative AI and the recommendation to consider generated texts without blindly relying on them. This distinction underscores the importance of maintaining critical thinking and human judgment in the face of AI-generated content or recommendations.

The court's careful distinction between 'considering' and 'relying' on COMPAS outputs serves as a model for how we might approach the integration of AI systems in various domains. It suggests a balanced approach that leverages the analytical capabilities of AI while preserving the essential role of human discretion and expertise.

4. Conclusion

The COMPAS case serves as a compelling illustration of the complexities involved in assessing algorithmic fairness, particularly when deploying AI systems in high-stakes domains like criminal justice. Through an educational journey, we have deconstructed the arguments and analyses put forth by various stakeholders, revealing the limitations and potential pitfalls of relying on any single fairness metric or analytical approach in isolation.

The case underscores the importance of adopting a nuanced, intersectional perspective that considers multiple dimensions of identity and their potential interactions. By examining gender disparities in addition to racial bias, we uncover potential sources of discrimination that may be obscured when focusing solely on a single axis of analysis.

Moreover, this case highlights the challenges of simultaneously satisfying different notions of fairness [14], a phenomenon known as the impossibility of fairness. It illustrates the trade-offs that must be navigated between statistical accuracy, social justice, and ethical considerations in the design and deployment of AI systems.

Ultimately, the COMPAS controversy serves as a cautionary tale against oversimplifying the complex issue of algorithmic fairness. It emphasizes the need for a holistic approach that considers the societal context, the potential for intersectional biases, and the inherent limitations of any single analytical framework.

References

[1] J. Dressel, H. Farid, The accuracy, fairness, and limits of predicting recidivism, Science advances 4 (2018) eaao5580.

[2] C. Rudin, C. Wang, B. Coker, The age of secrecy and unfairness in recidivism prediction, Harvard Data Science Review 2 (2020) 1.

[3] A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, Big data 5 (2017) 153–163.
[4] C. Engel, L. Linhardt, M. Schubert, Code is law: how compas affects the way the judiciary handles the risk of recidivism, Artificial Intelligence and Law (2024) 1–23.
[5] R. Berk, H. Heidari, S. Jabbari, M. Kearns, A. Roth, Fairness in criminal justice risk assessments: The state of the art, Sociological Methods & Research 50 (2021) 3–44.
[6] S. Badaloni, A. Rodà, et al., Gender knowledge and artificial intelligence, in: Proceedings of the 1st Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming, BEWARE-22, co-located with AIxIA, 2022.
[7] T. Brennan, W. Dieterich, B. Ehret, Evaluating the predictive validity of the compas risk and needs assessment system, Criminal Justice and behavior 36 (2009) 21–40.
[8] Northpointe, Compas risk & need assessment system: Selected questions posed by inquiring agencies, 2012.
[9] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, in: Ethics of data and analytics, Auerbach Publications, 2022, pp. 254–264.
[10] W. Dieterich, C. Mendoza, T. Brennan, Compas risk scales: Demonstrating accuracy equity and predictive parity, Northpointe Inc 7 (2016) 1–36.
[11] S. v. Loomis, Wisconsin supreme court requires warning before use of algorithmic risk assessments in sentencing, Harvard Law Review 130 (2017) 1530–1537.
[12] M. L. Cummings, Automation bias in intelligent time critical decision support systems, in: Decision making in aviation, Routledge, 2017, pp. 289–294.
[13] G. Tamburrini, The heuristics gap in ai ethics: Impact on green ai policies and beyond, Journal of Responsible Technology 21 (2025) 100104.
[14] F. Lagioia, R. Rovatti, G. Sartor, Algorithmic fairness through group parities? the case of compas-sapmoc, AI & SOCIETY 38 (2023) 459–478.