Utrecht University

Section 1: Intro

Assessing fairness measures for automatic decision-making models in the mental health domain

Informed Consent Agreement

Before participating in this study, please read the following information and indicate below whether you consent to the conditions of participation. As part of a research project focusing on the **fairness of AI models in the mental health domain**, we would like to understand the **clinicians' perspective on fairness**. Based on this input, we want to explore how these systems can be improved. For this purpose, we are asking you to contribute your perspective through participating in a survey. Your input will be used to gain more insight to inform future research that aims to improve AI-based decision-making systems in the mental health domain. Participation is completely voluntary. You can decide to withdraw at any time and for any reason, including after participating.

The survey will take approximately **10 minutes** and will be in the form of closed questions with optional comment fields. We intend to report the results of this survey in publications and/or presentations. Any information from the optional comment fields that could identify you as an individual, will be replaced or removed before publishing the survey data. Therefore, your identity as a participant will always remain confidential, and the risk of participation will be minimal. This project falls under fundamental research without any commercial purpose nor external stakeholders or partners.

For more information regarding this project, please contact:
Gizem Sogancioglu (PhD Student)
Social and Affective Computing Group, Utrecht University
g.sogancioglu@uu.nl

Dr. Pablo Mosteiro Romero (Assistant Professor)
Natural Language and Text Processing (NLTP) group, Utrecht University
p.j.mosteiroromero@uu.nl

*** Agreement**

if you confirm all the statements in the following paragraph, please agree the participation, and start the survey.

- ◯ I have read the information above, and understand the nature and goal of this research project. I understand my participation is voluntary, and that I may withdraw from the study at any time. I understand that any information I enter may be shared with the broader research community, and may be reported in scientific publications. Any information that could identify me as an individual will be replaced or removed. I agree to participate in the research project as described above.

Demographic Information

Background Information

The following information will be used to describe your background and general demographics.

What is your full name? (Optional)

* What is your profession?

* What is your background?

- psychiatry
- psychology
- computer science
- other

* How many years of experience do you have in this field?

* Please specify the highest degree or level of school you have completed:

- Bachelor’s degree
- Master’s degree
- Professional or doctoral degree (JD, MD, PhD)
- Enter yourself if not listed here

How knowledgeable are you with Electronic Health Records (EHR) data?

- Extremely knowledgeable
- Very knowledgeable
- Moderately knowledgeable
- Slightly knowledgeable
- Not knowledgeable at all

How knowledgeable are you with AI-based automatic decision-making models?

- Extremely knowledgeable
- Very knowledgeable
- Moderately knowledgeable
- Slightly knowledgeable
- Not knowledgeable at all

* What are your thoughts on gender fairness within the mental health domain, and how important is it to guarantee gender fairness prior to deploying automated predictive models in the mental health domain?

Depression V2

<u>Scenario: Depression phenotype recognition from clinical notes.</u>

While Electronic Health Records (EHRs) contain medical billing codes that aim to represent the conditions and treatments patients may have, much of the information is only present in the patient notes. These notes were previously written by clinicians about patients admitted to the hospital for any reason (suicide, heart attack, etc.). Extracting information related to a patient's condition and treatment provides a good knowledge base about the patient, which yields better treatment later.

We are interested in extracting depression phenotypes from these notes. It is a task that requires predicting the presence of a depression phenotype in clinical notes. Clinicians read the clinical notes and annotate whether the phenotype occurs.

This is the guideline that clinicians used to annotate the depression phenotype: "**diagnosis of depression, prescription of antidepressant medications, or any mention of intentional drug overdose, suicide, or attempts at self-harm"**

See sample text below.

Sample Text	**Gender**	**Depression Phenotype**
Female patient came to the hospital exhibiting signs of severe emotional distress. During the consultation, **she** disclosed past instances of intentional drug overdose and expressed ongoing suicidal ideation.	Female	YES

* If you were a decision-maker who needed to annotate depression phenotypes based on only the textual notes of EHR, would the **gender** of the patient influence your annotation? Why?

<u>Automatic machine learning model to predict depression phenotypes from clinical notes.</u>

Depression phenotype annotation requires extensive time and effort by clinicians. We want to automate this work with a machine learning model (ML). This ML model automatically identifies and annotates the depression phenotype with good accuracy. The same type of input text (see above) is given to the model, and it makes predictions using text.

However, we need to ensure the model treats all **gender** groups fairly. The given scenario uses only two genders (female and male) for the remaining questions.

* Is it fair if "**gender**" is used by the automatic **depression phenotype recognition** model? Please indicate your reasoning in the textbox.

- ◯ Yes
- ◯ Maybe
- ◯ No
- ◯ I do not know

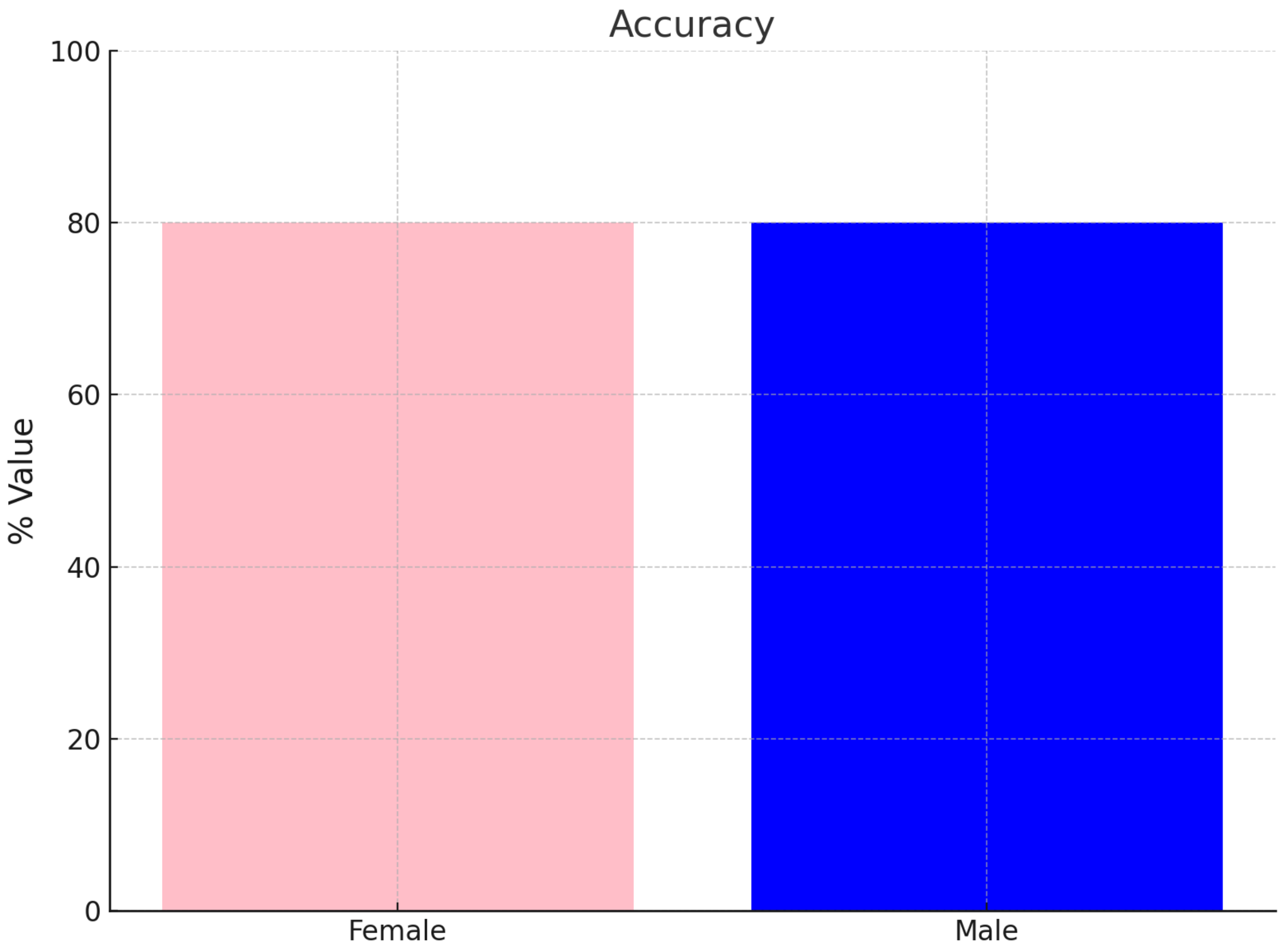
* How important do you think to ensure gender fairness in the context of recognizing depression phenotypes? When would you

consider the automatic algorithm fair for gender groups in a given problem?

* What is the more harmful error type for this problem between **False Positive** and **False Negative**?
False Negative: depression phenotype exists, but not found (under-treatment),
False Positive: depression phenotype does not exist, but was mistakenly found by the model and unnecessary intervention was taken (over-treatment)

- False Positive
- False Negative
- Equally harmful errors
- Other

Equal Accuracy:
This fairness measure ensures that a model is equally accurate for all groups, such as different genders. For this use-case, equal accuracy would mean that the tool is just as good at predicting depression and non-depression phenotypes accurately in men as it is in women. For example; the bar plot below shows that the model has 80% accuracy for both male and female examples; so the equal accuracy measure is satisfied.



* How clear is the **equal accuracy** measure?

- Extremely clear
- Very clear
- Somewhat clear
- Very unclear
- Extremely unclear

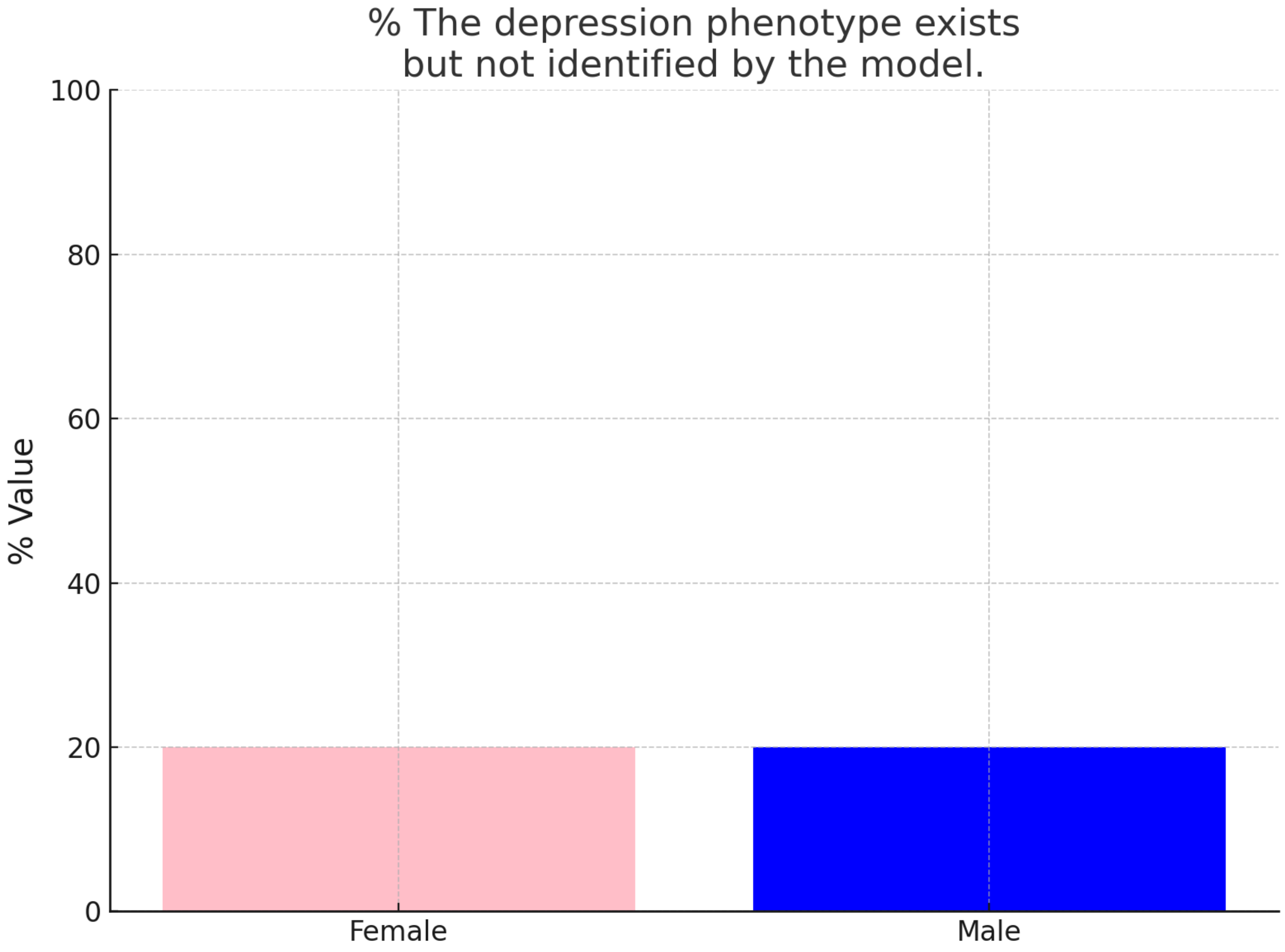
* How important is it to satisfy the **equal accuracy** measure for

depression phenotype recognition problem?

- Extremely important
- Very important
- Moderately important
- Slightly important
- Not at all important
- Other

Equal False Negative Rates:

This focuses on equalizing false negative rates across gender groups. A false negative happens when the system fails to identify a "depression phenotype" that is present in the text. For example; the bar plot below shows that the model misses depression phenotypes equally for female and male groups.



* How clear is the **equal false negative rates** measure?

- ◯ Extremely clear
- ◯ Very clear
- ◯ Somewhat clear
- ◯ Very unclear
- ◯ Extremely unclear

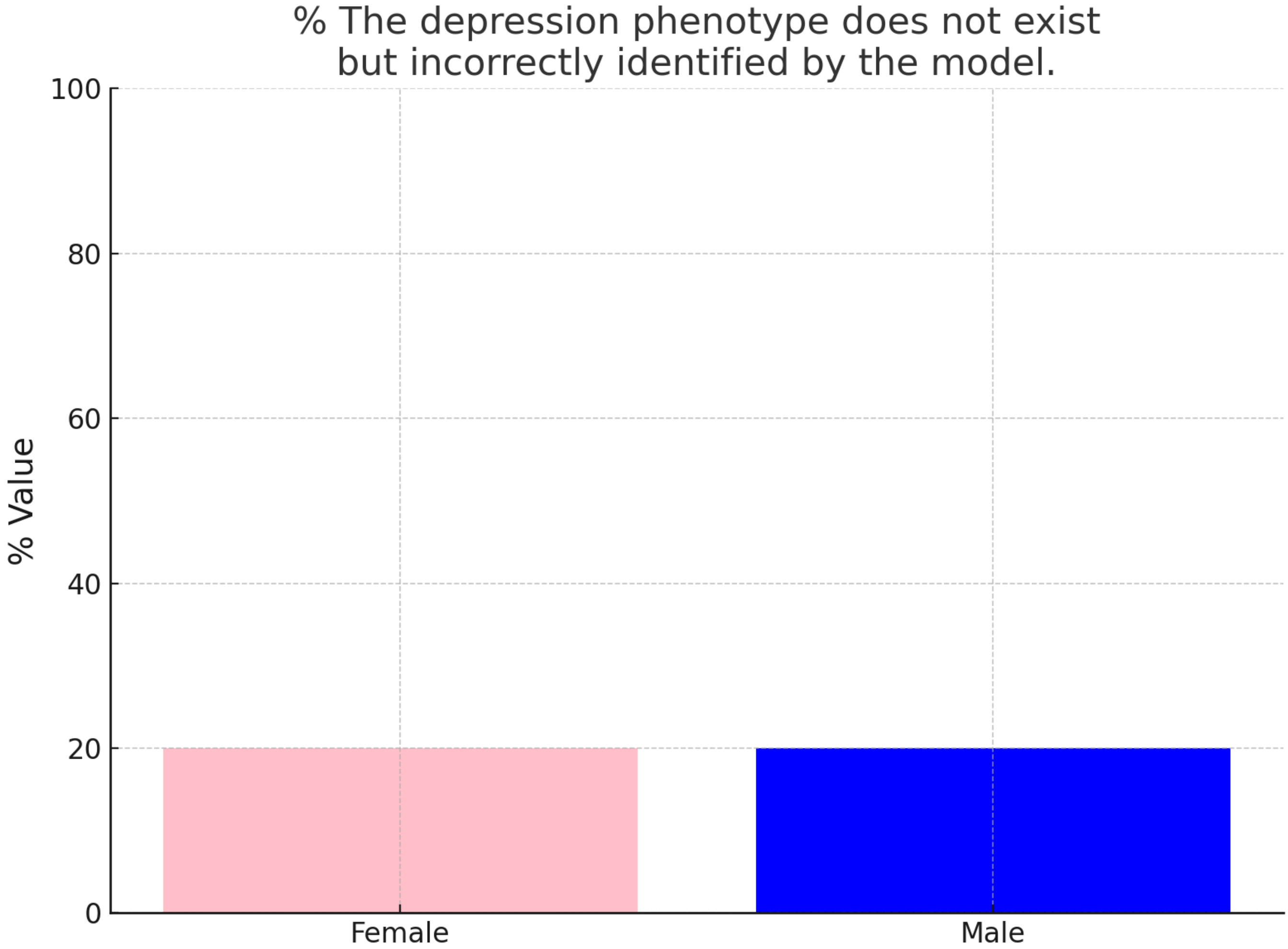
* How important is it to satisfy the **equal false negative**

rates measure for **depression phenotype recognition** problem?

- Extremely important
- Very important
- Moderately important
- Slightly important
- Not at all important
- Other

Equal False Positive Rates:

This focuses on equalizing false positive rates across gender groups. A false positive occurs when the system incorrectly predicts a "depression phenotype", while it is not present in the text. For example; the bar plot below shows that the model misidentifies depression phenotypes equally for female and male groups.



* How clear is the **equal false positive rates** measure?

- ◯ Extremely clear
- ◯ Very clear
- ◯ Somewhat clear
- ◯ Very unclear
- ◯ Extremely unclear

* How important is it to satisfy the **equal false positive**

rates measure for **depression phenotype recognition** problem?

- Extremely important
- Very important
- Moderately important
- Slightly important
- Not at all important
- Other

Counterfactual Token Fairness:

Let's assume that we have two notes as given in the table below. These notes are exactly the same except for the difference in gender pronouns. Counterfactual token fairness requires the model to make the same prediction for these note pairs.

Input 1 (she pronoun)	**Input 2 (he pronoun)**
Female patient came to the hospital exhibiting signs of severe emotional distress. During the consultation, **she** disclosed past instances of intentional drug overdose and expressed ongoing suicidal ideation.	**Male** patient came to the hospital exhibiting signs of severe emotional distress. During the consultation, **he** disclosed past instances of intentional drug overdose and expressed ongoing suicidal ideation.

* How clear is the **counterfactual token fairness** measure?

- Extremely clear
- Very clear
- Somewhat clear
- Very unclear
- Extremely unclear

*How important is it that **depression phenotype predictions** are similar for these two notes?

- Extremely important
- Very important
- Moderately important
- Slightly important
- Not at all important
- [ ] Other

* In practice, it is impossible to satisfy all fairness measures simultaneously. Please select the importance of the following measures for the **depression phenotype recognition** task. (must be satisfied > should be satisfied > maybe important > not very important > not relevant measure)

Reminder:

Equal Accuracy:	ensures equal accuracy across gender groups.
Equal False Negative Rates:	ensures equal false negative rates across gender groups.
Equal False Positive Rates:	ensures equal false positive rates across gender groups.
Counterfactual token fairness:	ensures the same prediction for all gender groups with identical clinical notes.

	must be satisfied	should be satisfied	maybe important	not very important	not relevant measure
Equal Accuracy	○	○	○	○	○
Equal False Negative Rates	○	○	○	○	○
Equal False Positive Rates	○	○	○	○	○
Counterfactual token fairness	○	○	○	○	○

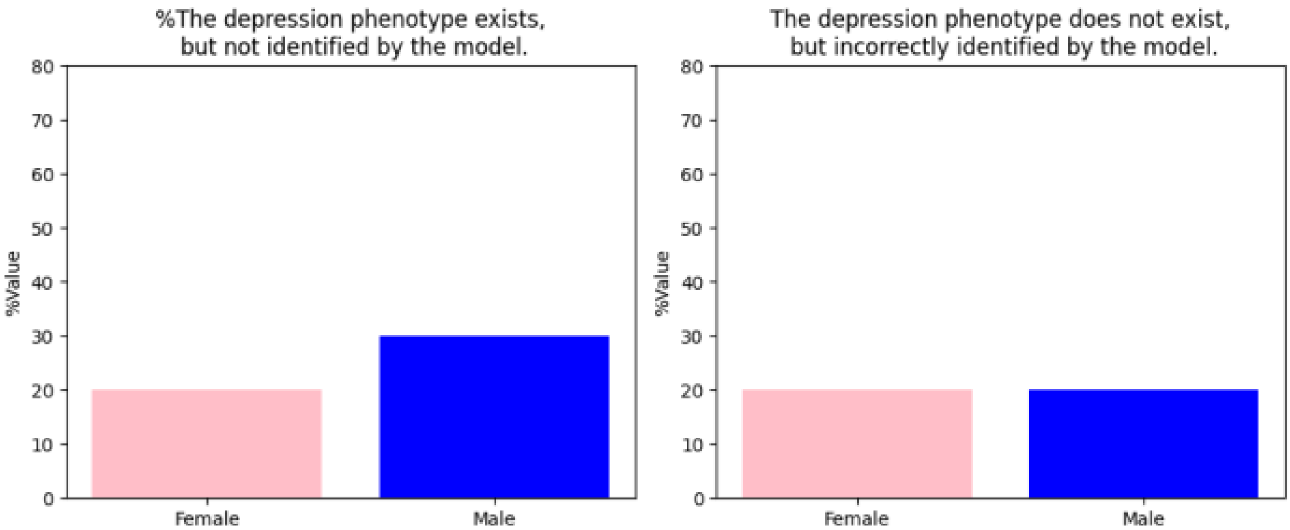
Pairwise Model Selection

You will see different model pairs below. Please select the one you think is fairer than the other.

* Which model is fairer, Model 1 or Model 2?

Model 1

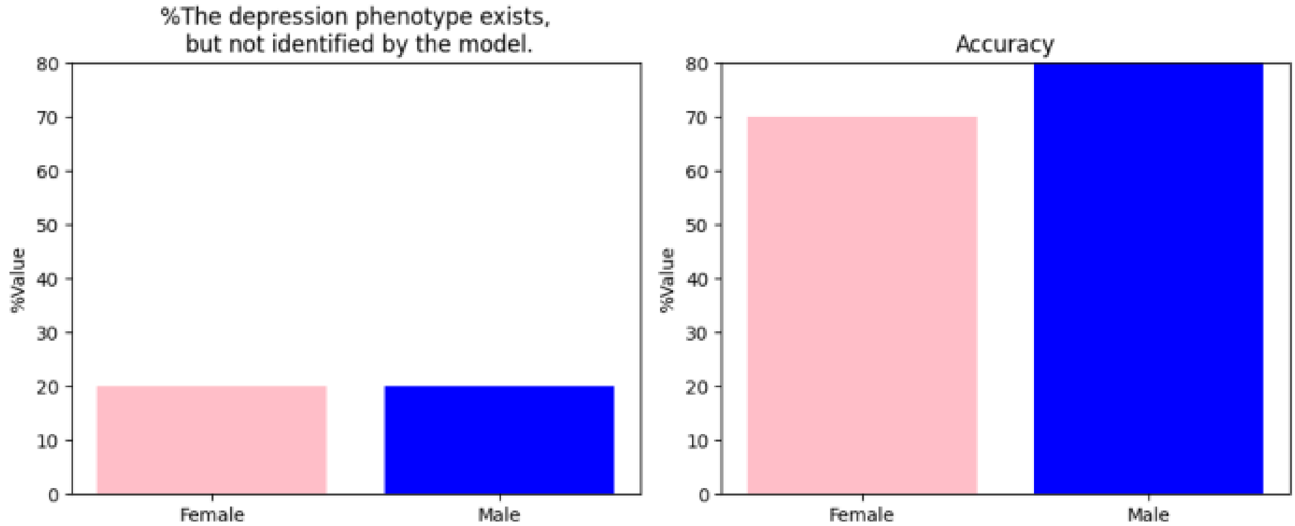
Model 2



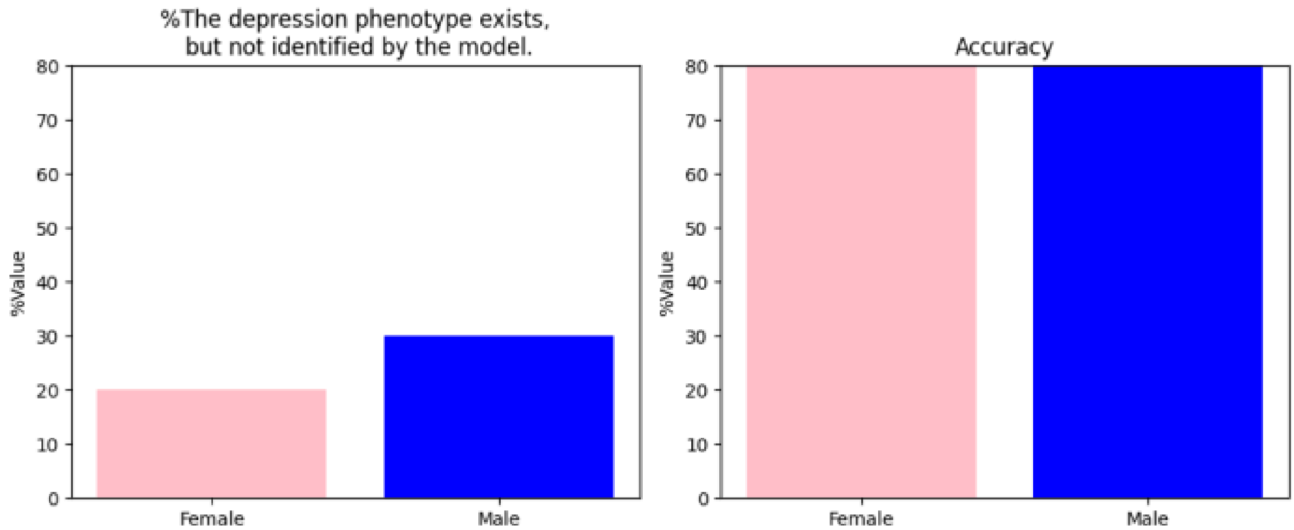
- Model 1
- Model 2
- Equally fair
- Equally unfair

* Which model is fairer, Model 1 or Model 2?

Model 1



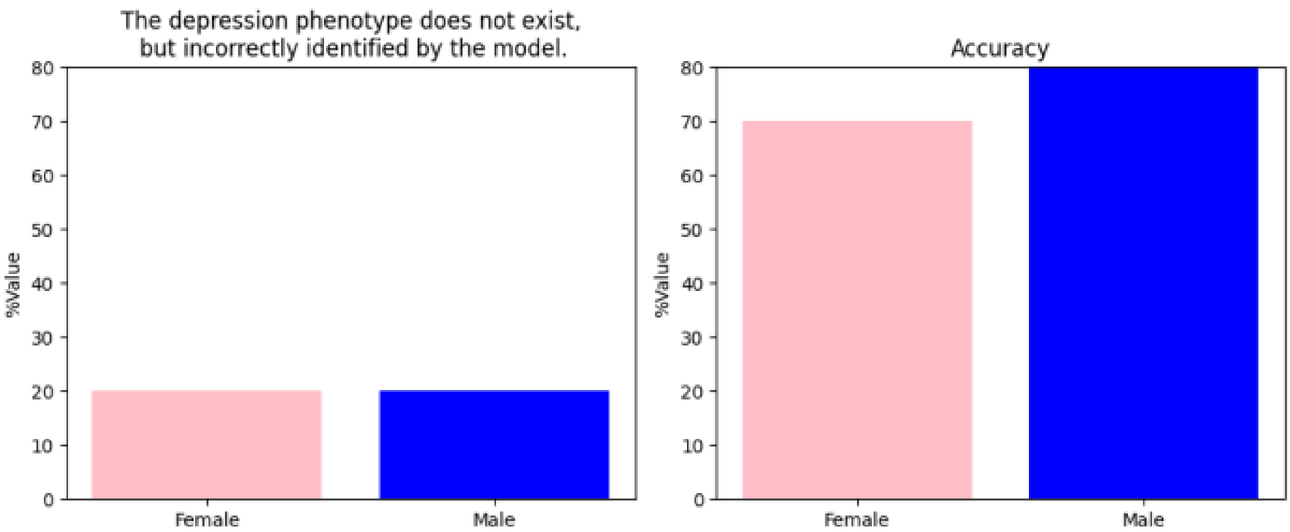
Model 2



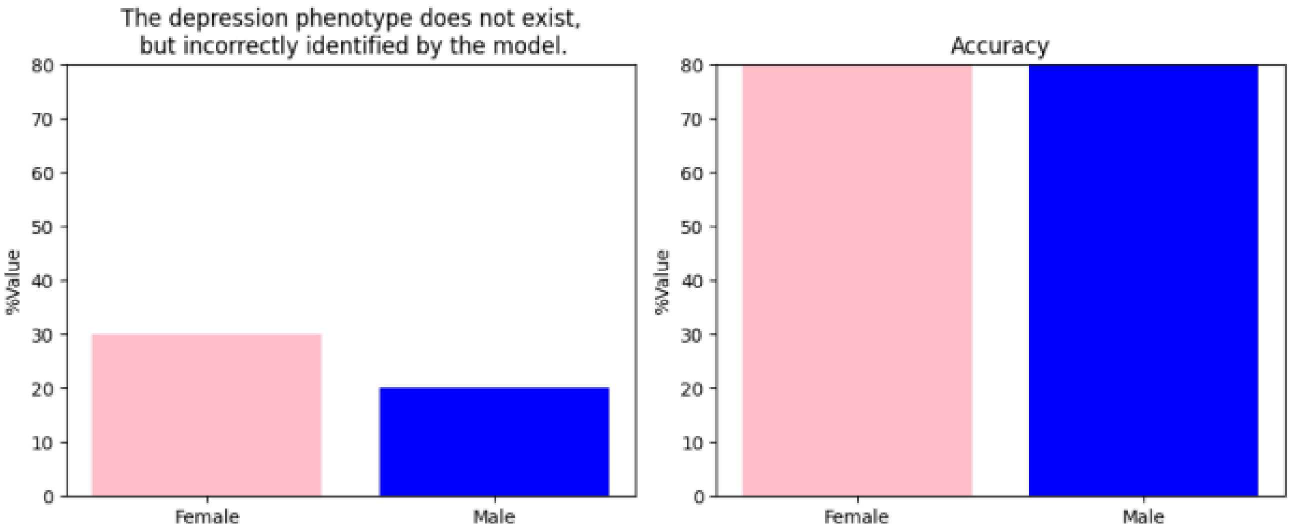
- ◯ Model 1
- ◯ Model 2
- ◯ Equally fair
- ◯ Equally unfair

* Which model is fairer, Model 1 or Model 2?

Model 1



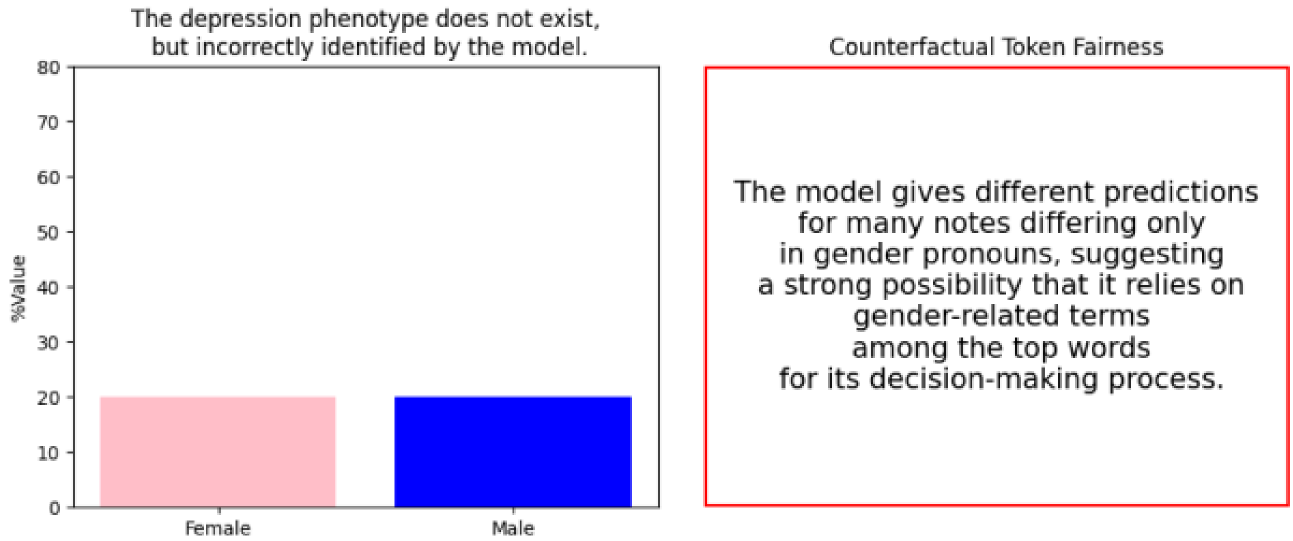
Model 2



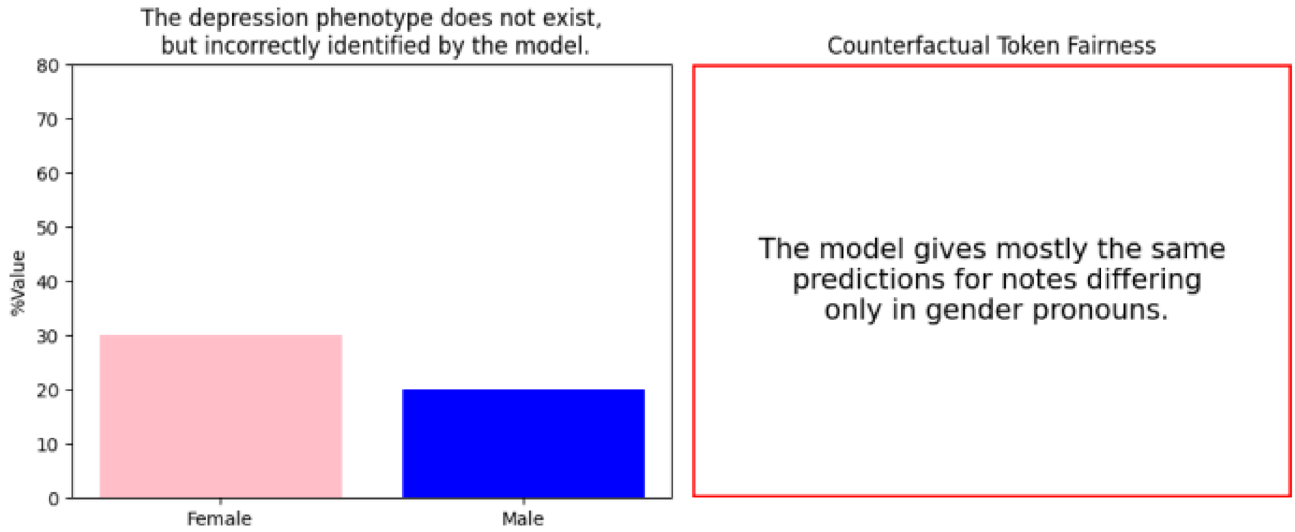
- ◯ Model 1
- ◯ Model 2
- ◯ Equally fair
- ◯ Equally unfair

* Which model is fairer, Model 1 or Model 2?

Model 1



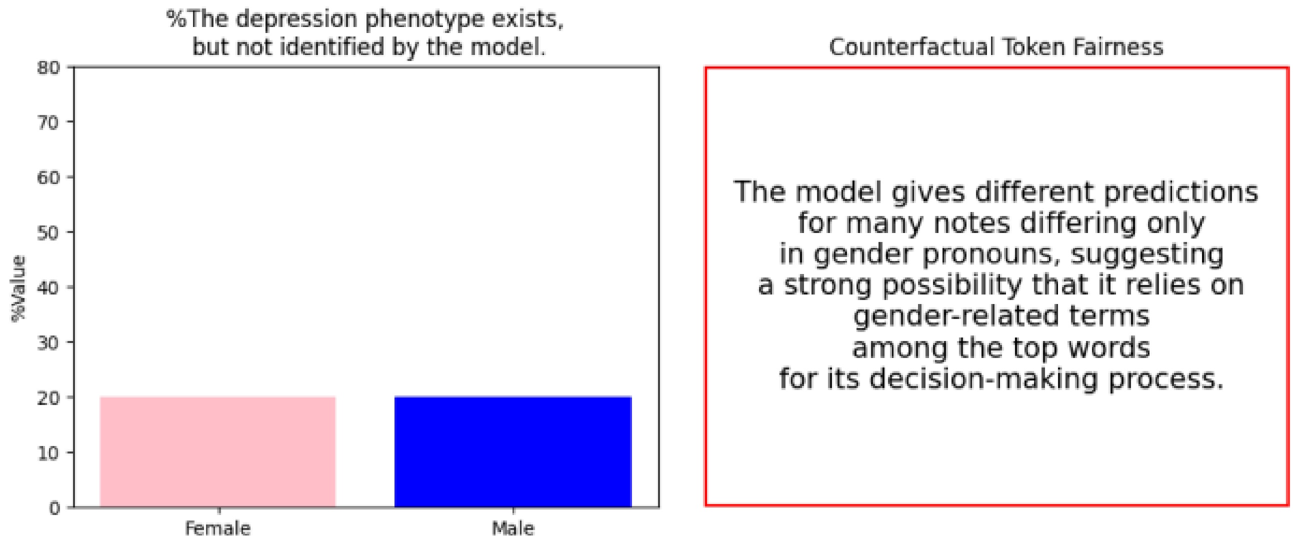
Model 2



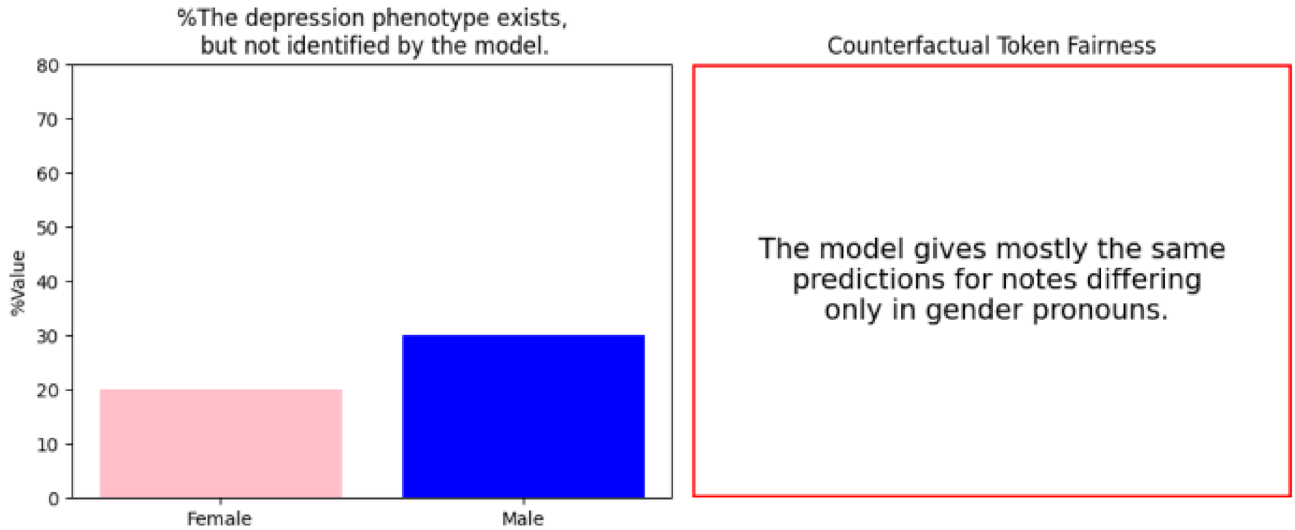
- Model 1
- Model 2
- Equally fair
- Equally unfair

* Which model is fairer, Model 1 or Model 2?

Model 1

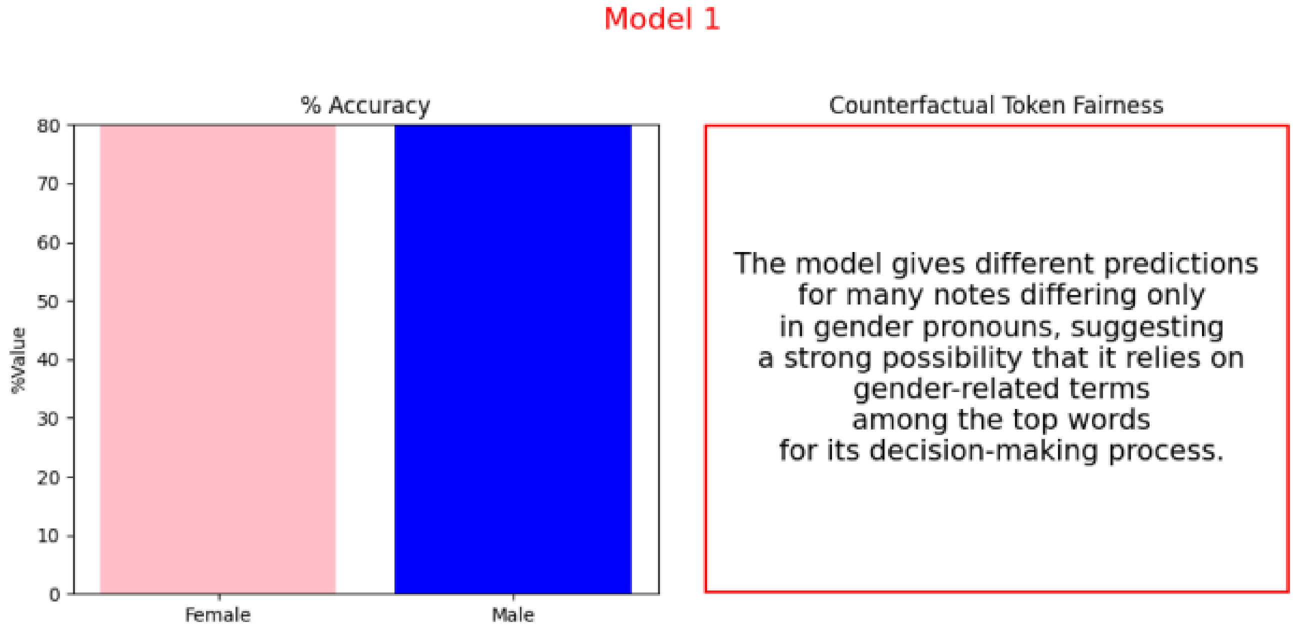


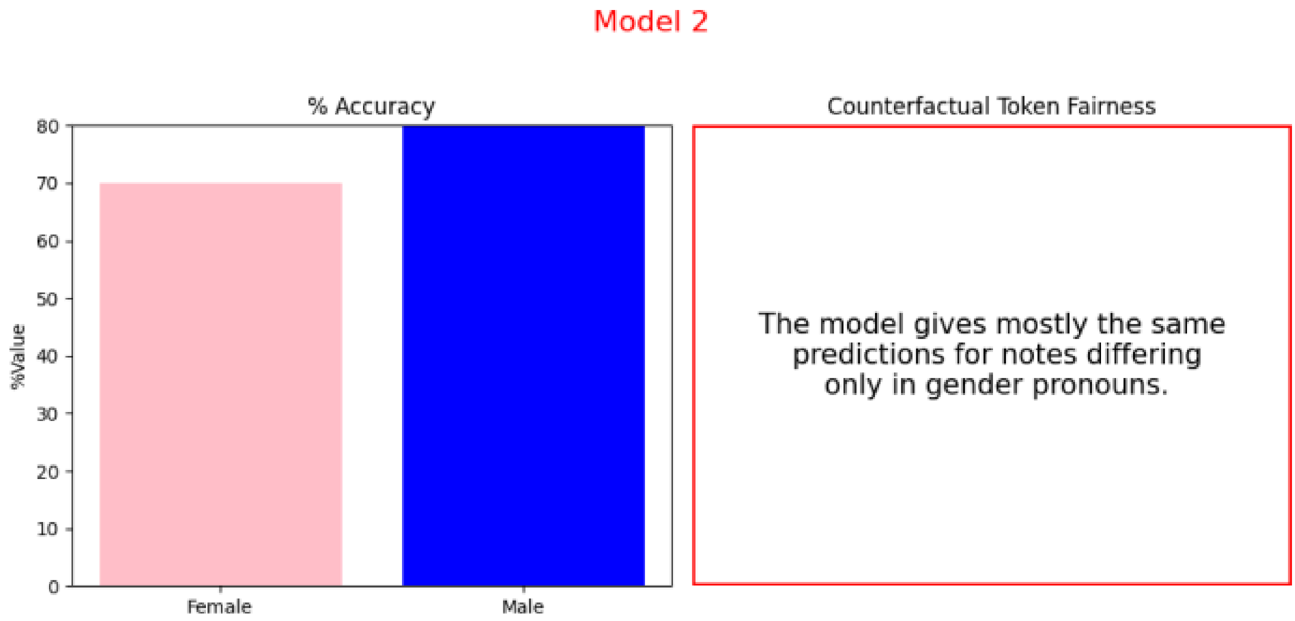
Model 2



- ◯ Model 1
- ◯ Model 2
- ◯ Equally fair
- ◯ Equally unfair

* Which model is fairer, Model 1 or Model 2?





- ◯ Model 1
- ◯ Model 2
- ◯ Equally fair
- ◯ Equally unfair

Violence V2

Scenario: Violence risk assessment from clinical notes.

There is a **machine learning model** that predicts whether a patient will exhibit **violence** and cause an **incident** in the first 28 days after the admission date, based on the clinical notes. This automatic decision-making model will replace the time-consuming questionnaire-based method (Brøset Violence Checklist). However, we need to ensure the model treats **all gender groups** fairly. The given scenario uses only two genders (**female** and **male**) for the remaining questions.

Example input text for the model:

Text	**Gender**	**Violence occurred**
[PERSON 1] comes in looking disheveled. Multiple cuts and bruises on arms. Behaving erratically. **He** was taken to **his** room. Complaints about headaches, given 400mg of ibuprofen for headache. The next day refuses to come out of **his** room.	Male	YES

* If you were a **decision-maker** who needed to predict whether a patient could have violent incidents, and you only had access to clinical textual notes in Electronic Health Records, would the **gender** of the patient influence your annotation? Why?

* Is it fair if "**gender**" is used by **violence risk prediction** model? Please indicate your reasoning in the textbox.

- Yes
- Maybe
- No
- I do not know

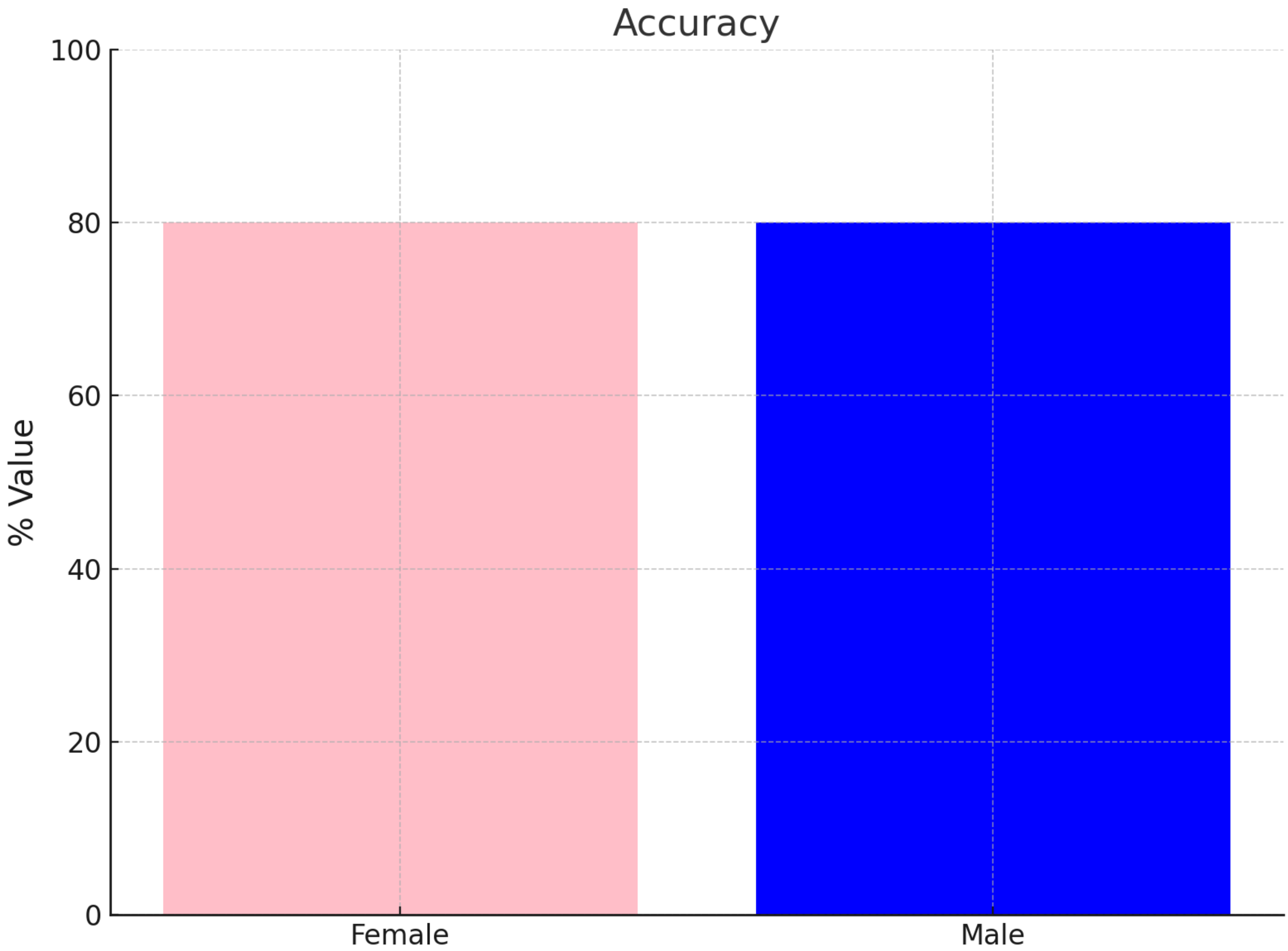
* How do you define gender fairness in the context of predicting violence incidents? When would you consider the automatic algorithm fair for gender groups in a given problem?

* What is the more harmful error type for this problem between False Positive and False Negative?

False Negative: violence occurred, but no intervention was taken (under-treatment)
False Positive: violence did not occur, but unnecessary intervention was taken (over-treatment)

- False Negative
- False Positive
- Equally harmful errors
- Add your answer

<u>Equal Accuracy:</u>
As explained before, this fairness measure ensures that a model is equally accurate for all groups, such as different genders.
For this use-case, equal accuracy would mean that the tool is just as good at predicting violent and non-violent incidents accurately in men as it is in women. For example; the bar plot below shows that the model has 80% accuracy for both male and female examples; so the equal accuracy measure is satisfied.



* How clear is the **equal accuracy** measure?

- Extremely clear
- Very clear
- Somewhat clear
- Very unclear
- Extremely unclear

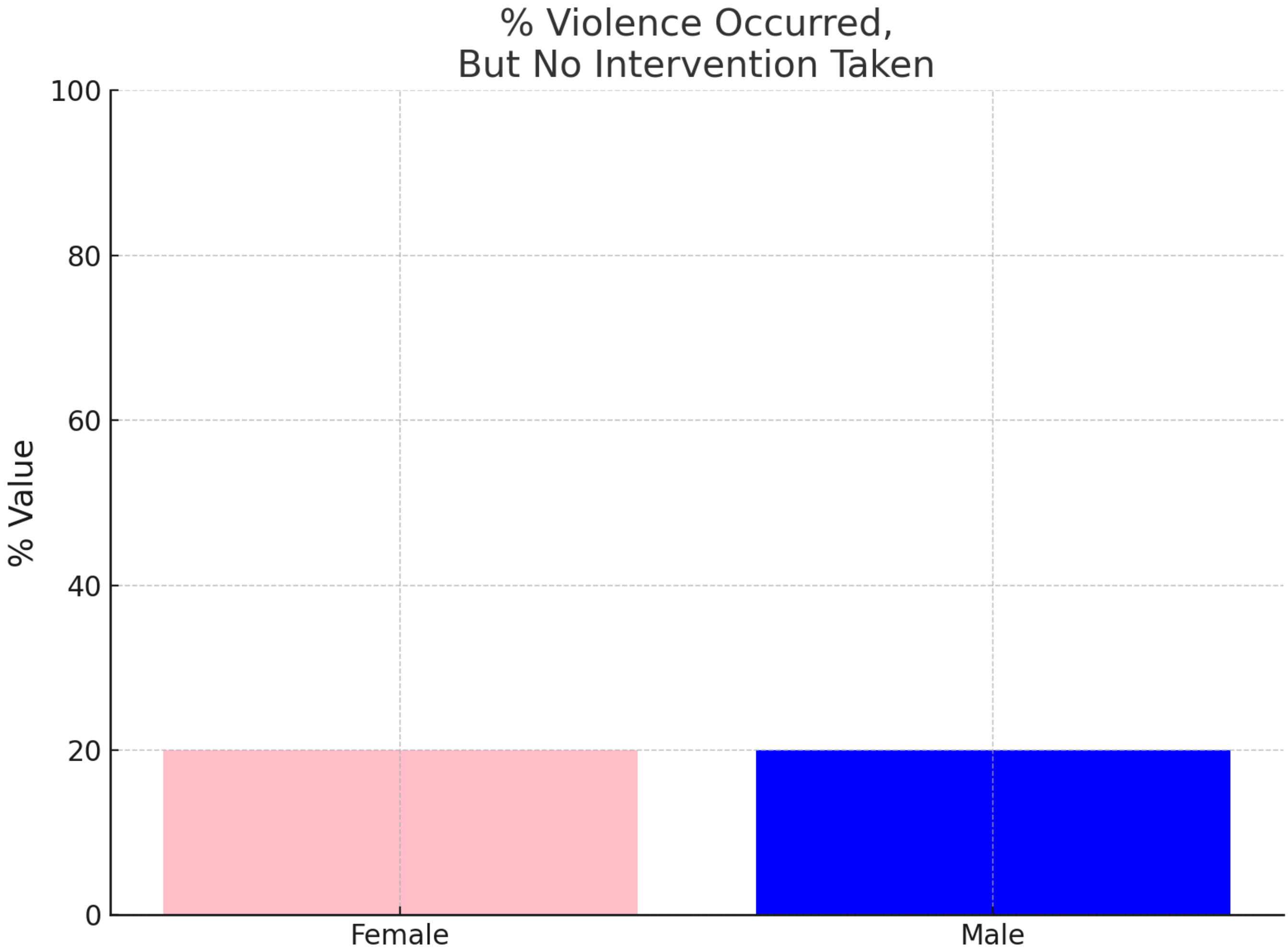
* How important is it to satisfy the **equal accuracy** measure for

violence risk prediction problem?

- Extremely important
- Very important
- Moderately important
- Slightly important
- Not at all important
- Other

Equal False Negative Rates:

As explained before, this focuses on equalizing false negative rates across gender groups.
For this use-case; a false negative occurs when the system predicts "no violence", but a violent incident does occur. It ensures that the model is equally accurate for all genders, particularly in correctly identifying actual violence cases. For example; the bar plot below shows that the model misses violence cases equally for female and male groups.



* How clear is the **equal false negative rates** measure?

- Extremely clear
- Very clear
- Somewhat clear
- Very unclear
- Extremely unclear

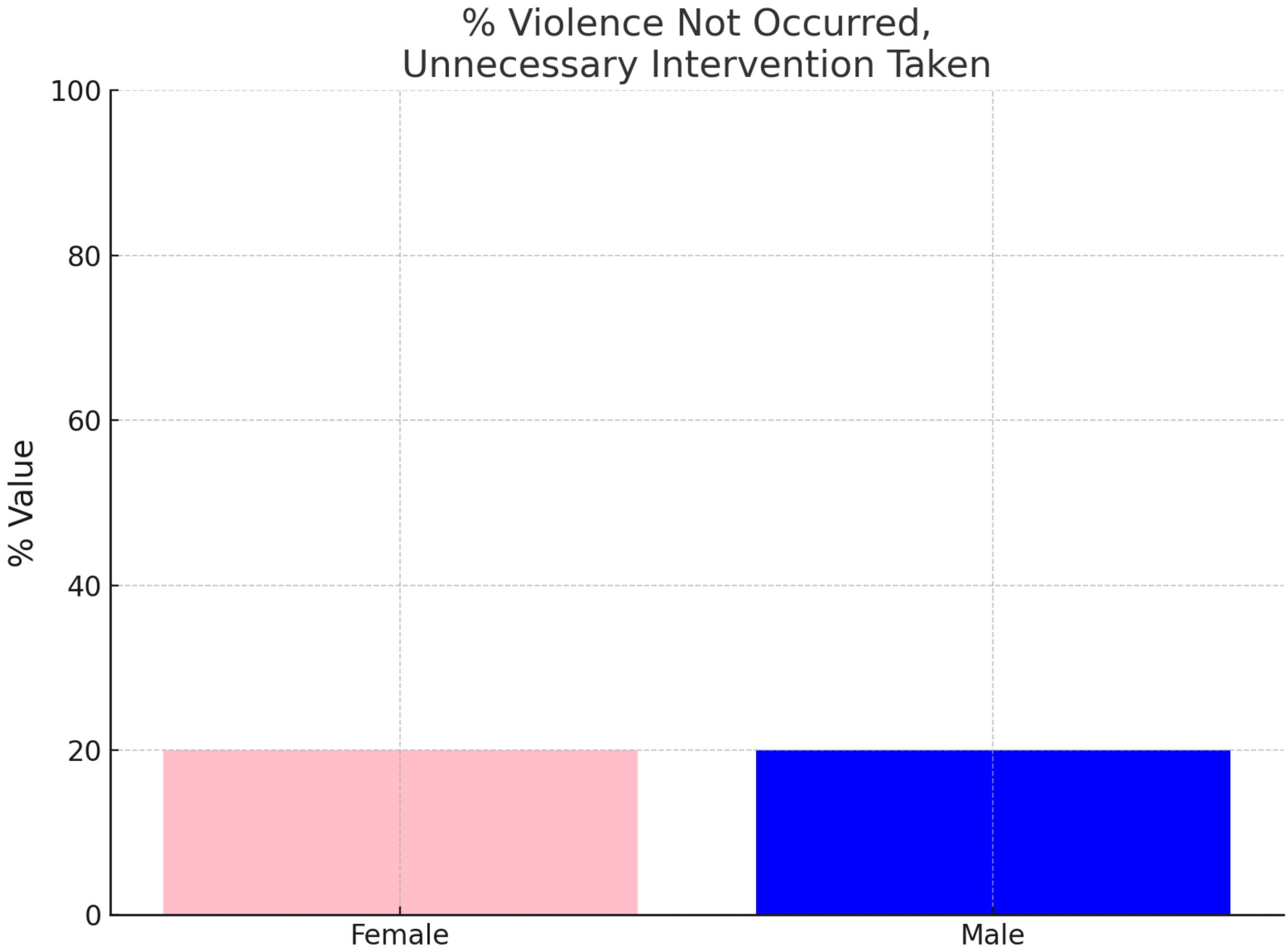
* How important is it to satisfy the **equal false negative**

rates measure for **violence risk prediction** problem?

- ◯ Extremely important
- ◯ Very important
- ◯ Moderately important
- ◯ Slightly important
- ◯ Not at all important
- ◯ [ ] Other

Equal False Positive Rates:

As explained before, this focuses on equalizing false positive rates across gender groups.
For this use-case, a false positive occurs when the system predicts "violence", but no violent incident occurs. It ensures that the model is equally accurate for all genders, particularly in correctly identifying actual non-violent cases. For example; the bar plot below shows that the model identifies non-violent cases equally for female and male groups.



* How clear is the **equal false positive rates** measure?

- Extremely clear
- Very clear
- Somewhat clear
- Very unclear
- Extremely unclear

* How important is it to satisfy the **equal false positive**

rates measure for **violence risk prediction** problem?

- ◯ Extremely important
- ◯ Very important
- ◯ Moderately important
- ◯ Slightly important
- ◯ Not at all important
- ◯ [ ] Other

Counterfactual Token Fairness:

Let's assume that we have two notes as given in the table below. These notes are exactly the same except for the difference in gender pronouns. Counterfactual token fairness requires the model to make the same prediction for these note pairs.

Input 1 (he pronoun)	**Input 2 (she pronoun)**
[PERSON 1] comes in looking disheveled. Multiple cuts and bruises on arms. Behaving erratically. **He** was taken to **his** room. Complaints about headaches, given 400mg of ibuprofen for headache. The next day refuses to come out of **his** room.	[PERSON 1] comes in looking disheveled. Multiple cuts and bruises on arms. Behaving erratically. **She** was taken to **her** room. Complaints about headaches, given 400mg of ibuprofen for headache. The next day refuses to come out of **her** room.

* How clear is the **<u>counterfactual token fairness</u>** measure?

- ◯ Extremely clear
- ◯ Very clear
- ◯ Somewhat clear
- ◯ Very unclear
- ◯ Extremely unclear

*How important is it that **violence risk prediction** decisions are similar for these two notes?

- ◯ Extremely important
- ◯ Very important
- ◯ Moderately important
- ◯ Slightly important
- ◯ Not at all important
- ◯ [ ] Other

* In practice, it is impossible to satisfy all fairness measures simultaneously. Please select the importance of the following measures for the **violence risk prediction** task. (must be satisfied > should be satisfied > maybe important > not very important > not relevant measure)

Reminder:

Equal Accuracy:	ensures equal accuracy across gender groups.
Equal False Negative Rates:	ensures equal false negative rates across gender groups.
Equal False Positive Rates:	ensures equal false positive rates across gender groups.
Counterfactual token fairness:	ensures the same prediction for all gender groups with identical clinical notes.

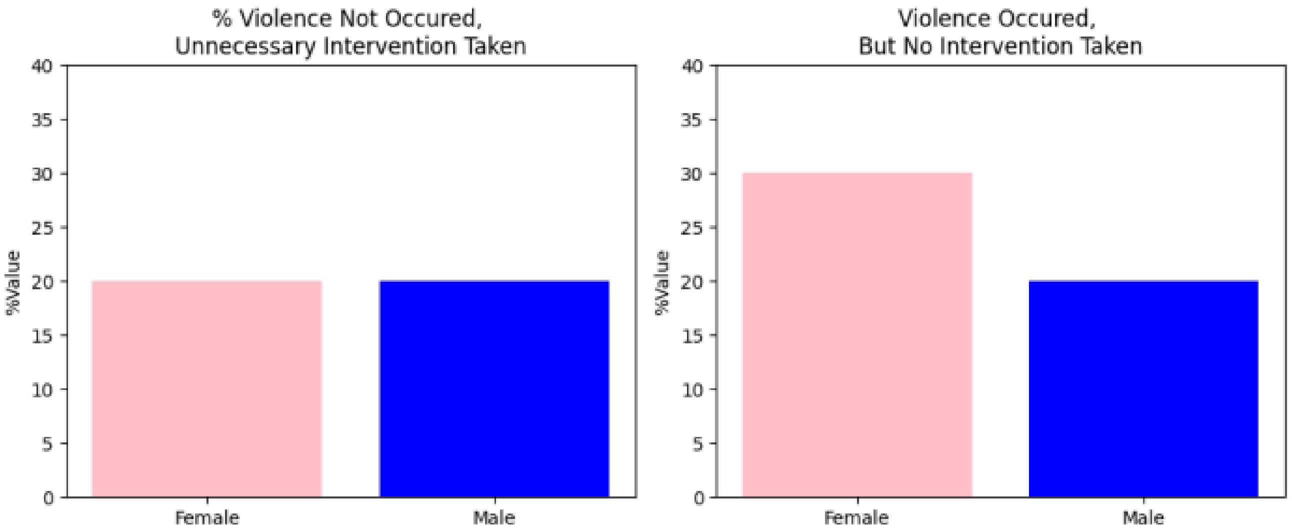
	must be satisfied	should be satisfied	maybe important	not very important	not relevant measure
Equal Accuracy	○	○	○	○	○
Equal False Negative Rates	○	○	○	○	○
Equal False Positive Rates	○	○	○	○	○
Counterfactual token fairness	○	○	○	○	○

Pairwise Model Selection

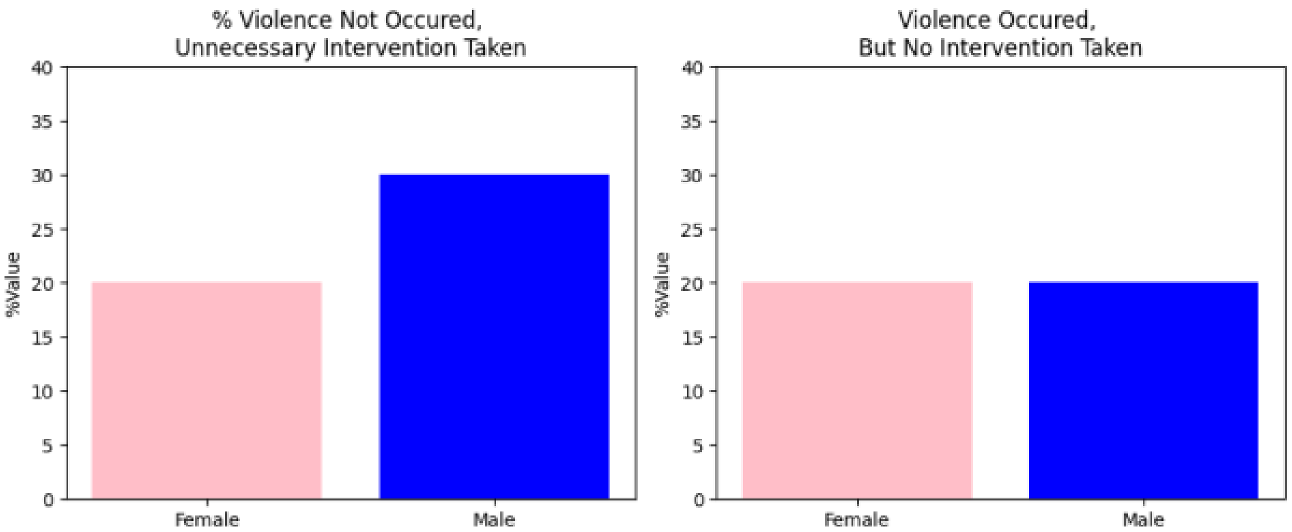
You will see different model pairs below. Please select the one you think is fairer than the other.

* Which model is fairer; Model 1 or Model 2?

Model 1



Model 2

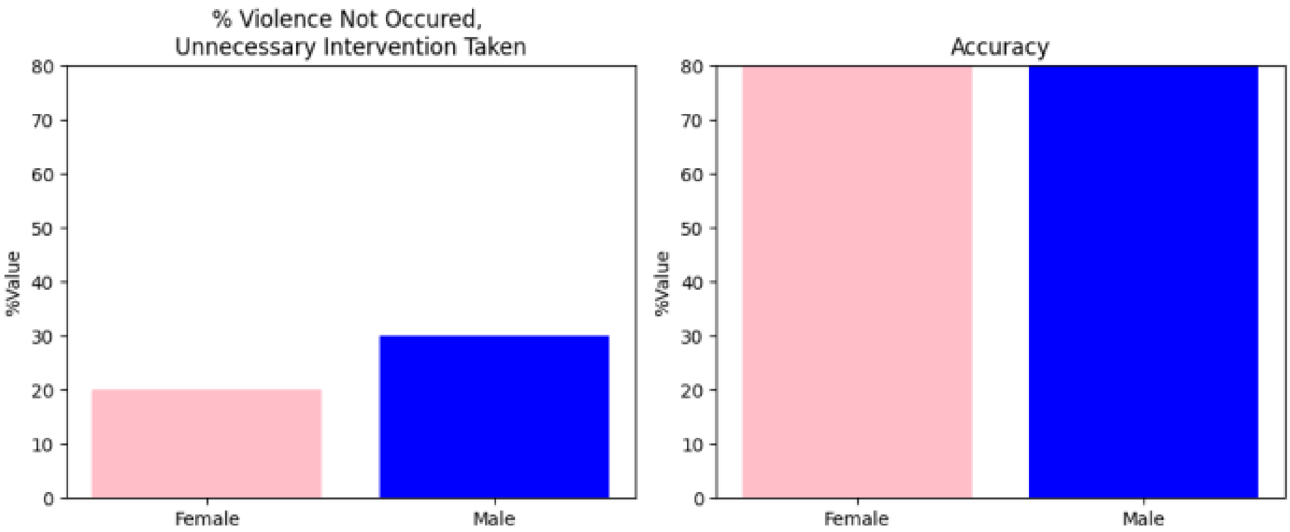


- Model 1
- Model 2
- Equally fair

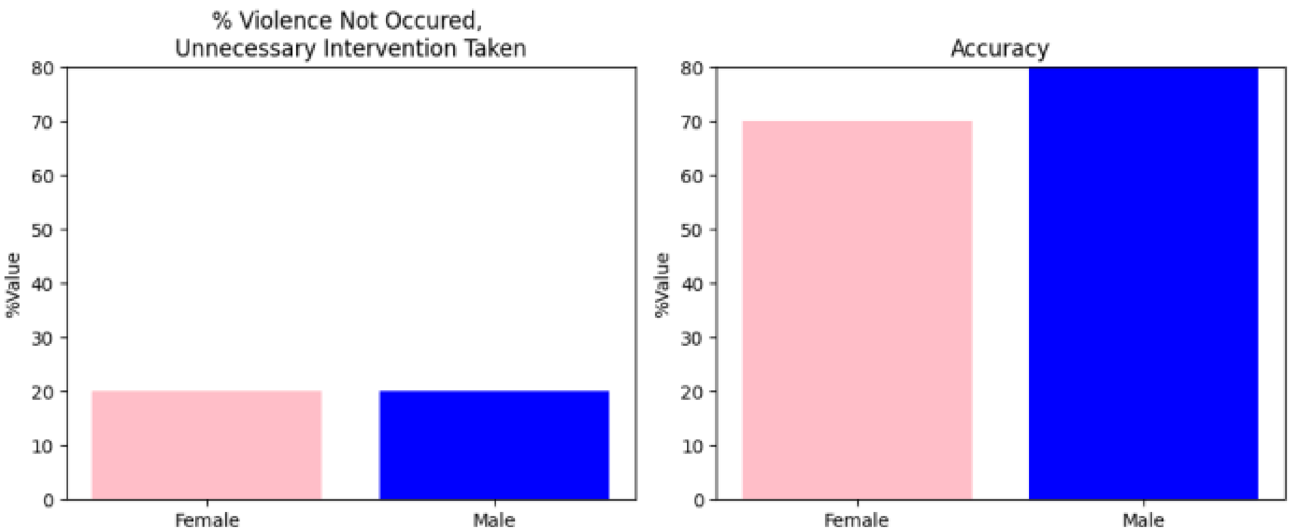
○ Equally unfair

* Which model is fairer; Model 1 or Model 2?

Model 1



Model 2



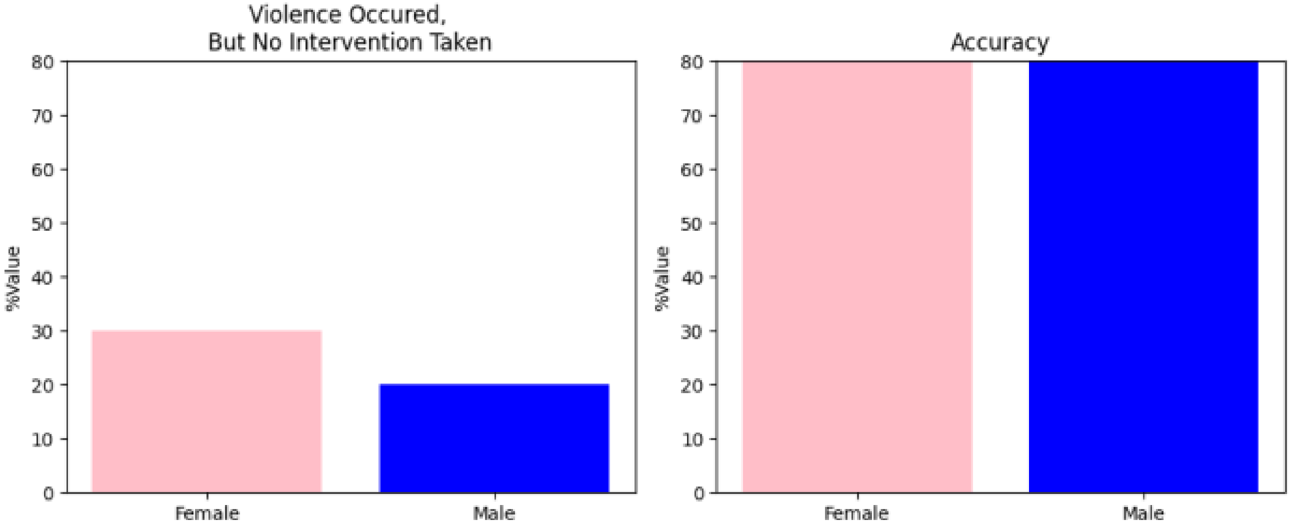
○ Model 1

○ Model 2

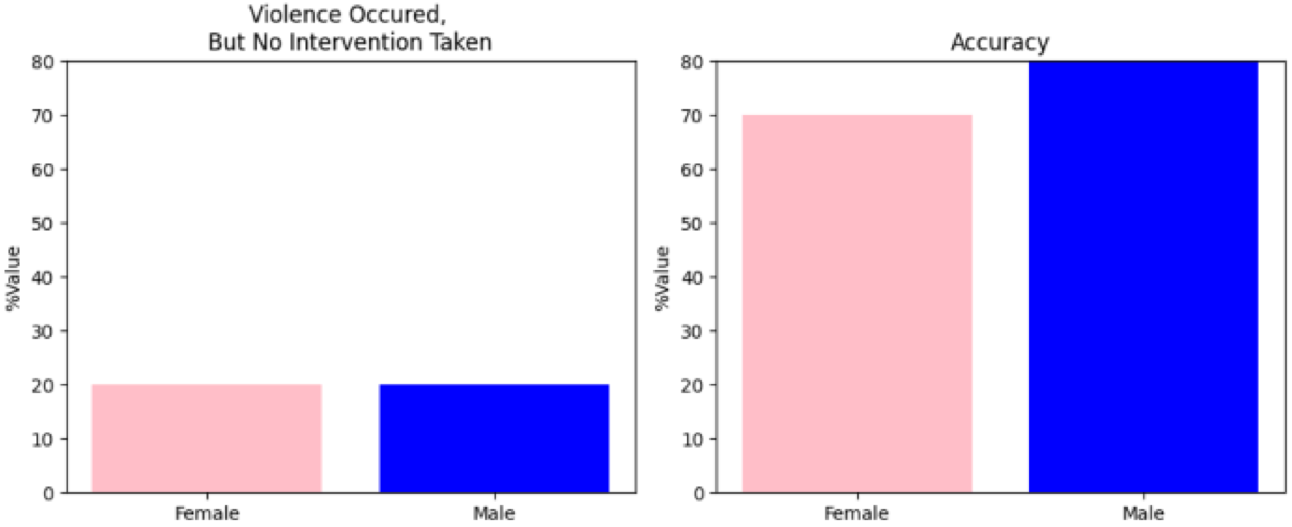
- ◯ Equally fair
- ◯ Equally unfair

* Which model is fairer; Model 1 or Model 2?

Model 1



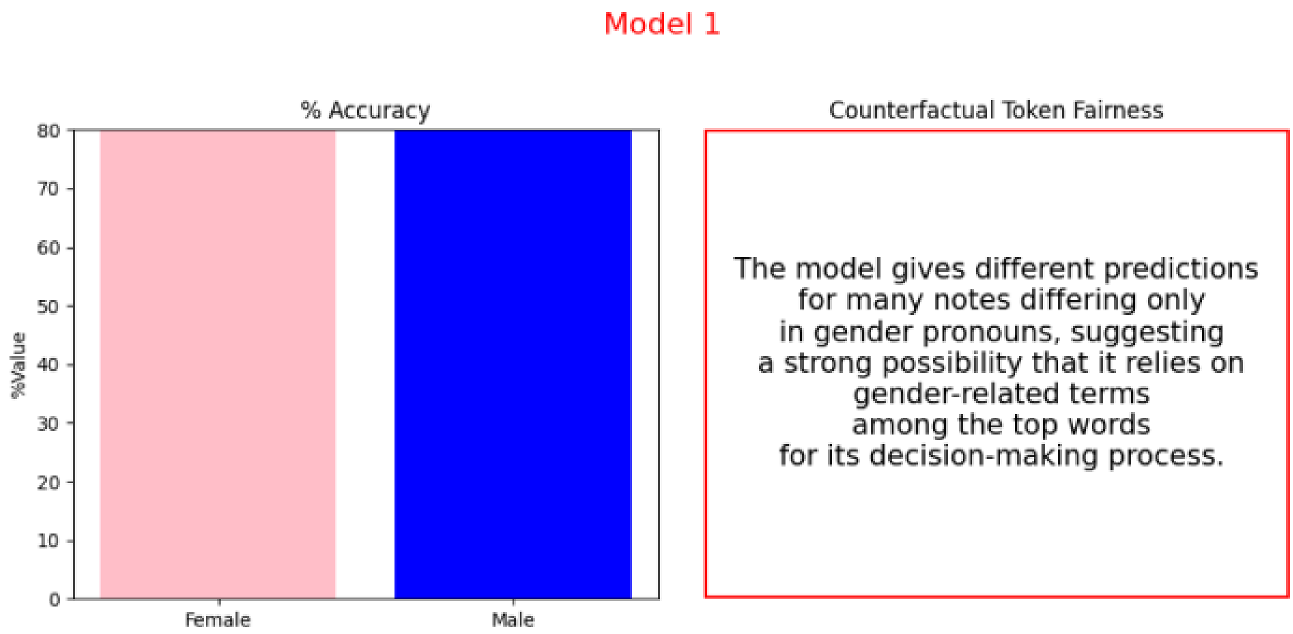
Model 2

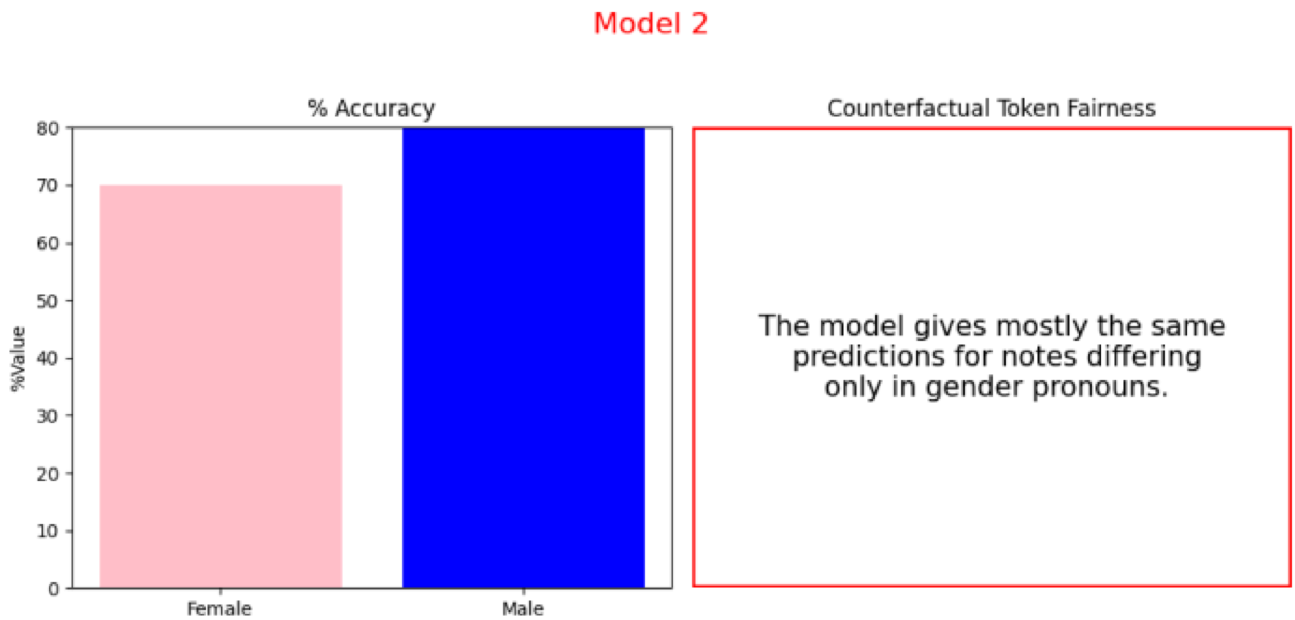


- ◯ Model 1

◯ Model 2
◯ Equally fair
◯ Equally unfair

* Which model is fairer; Model 1 or Model 2?

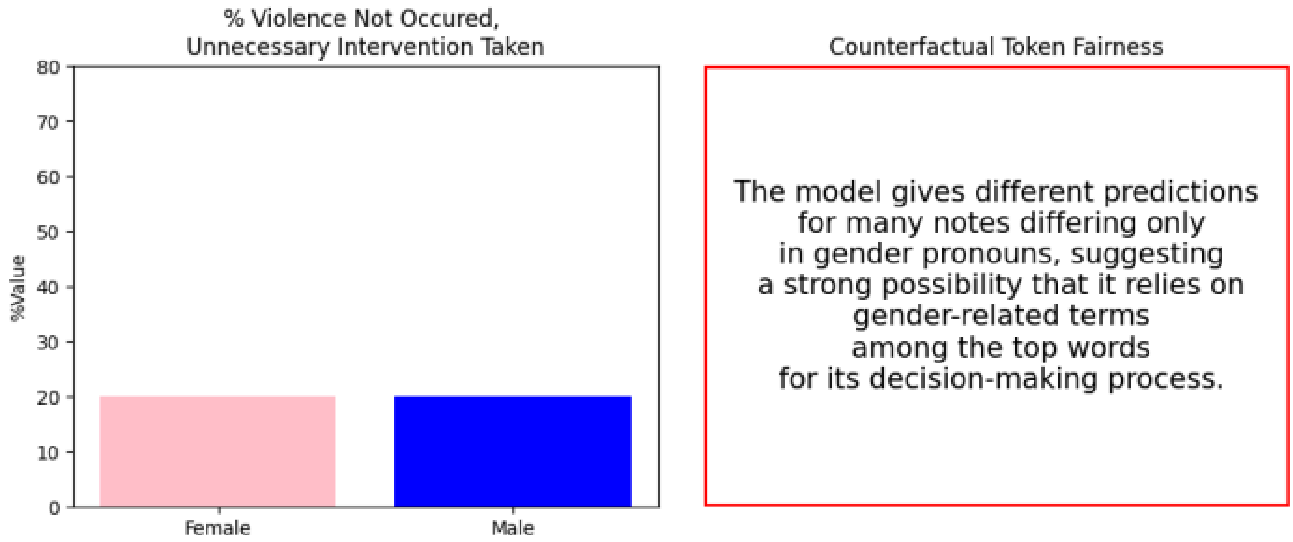




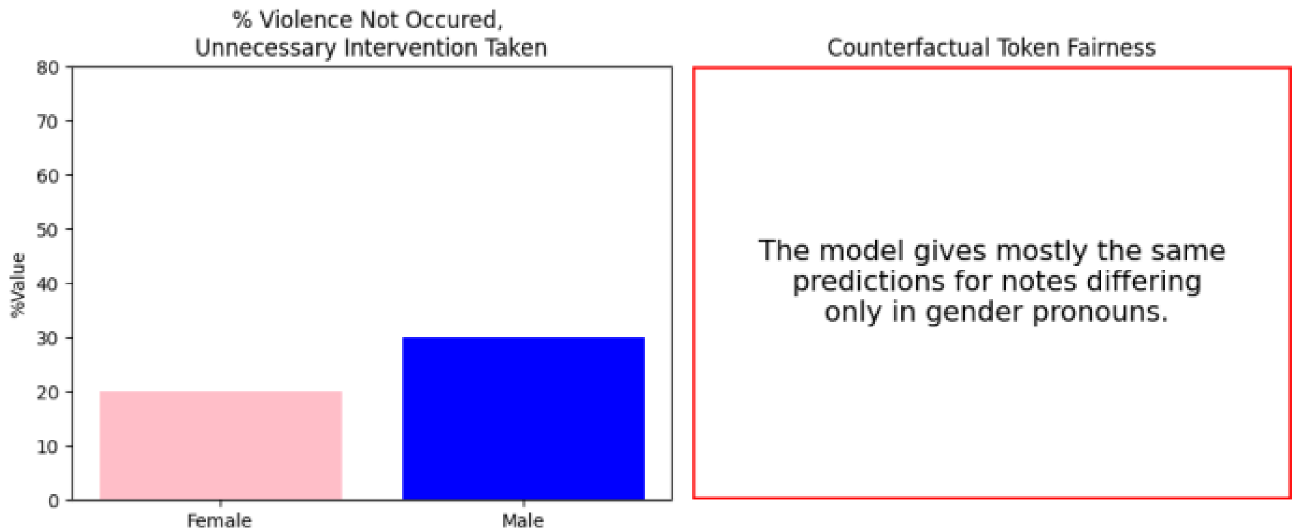
- ◯ Model 1
- ◯ Model 2
- ◯ Equally fair
- ◯ Equally unfair

* Which model is fairer; Model 1 or Model 2?

Model 1



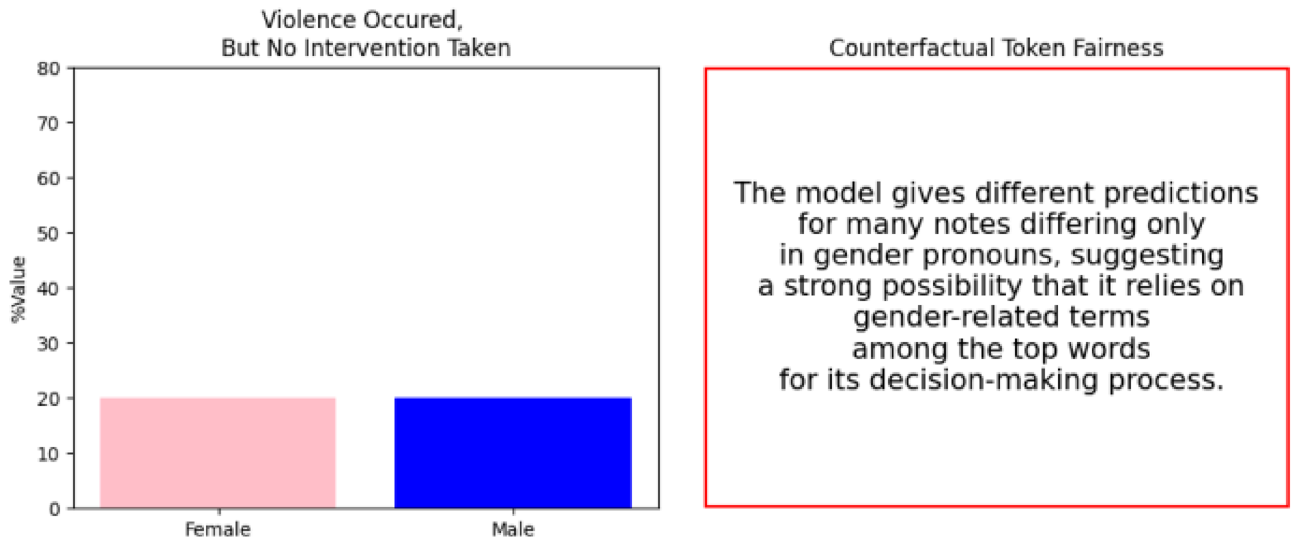
Model 2



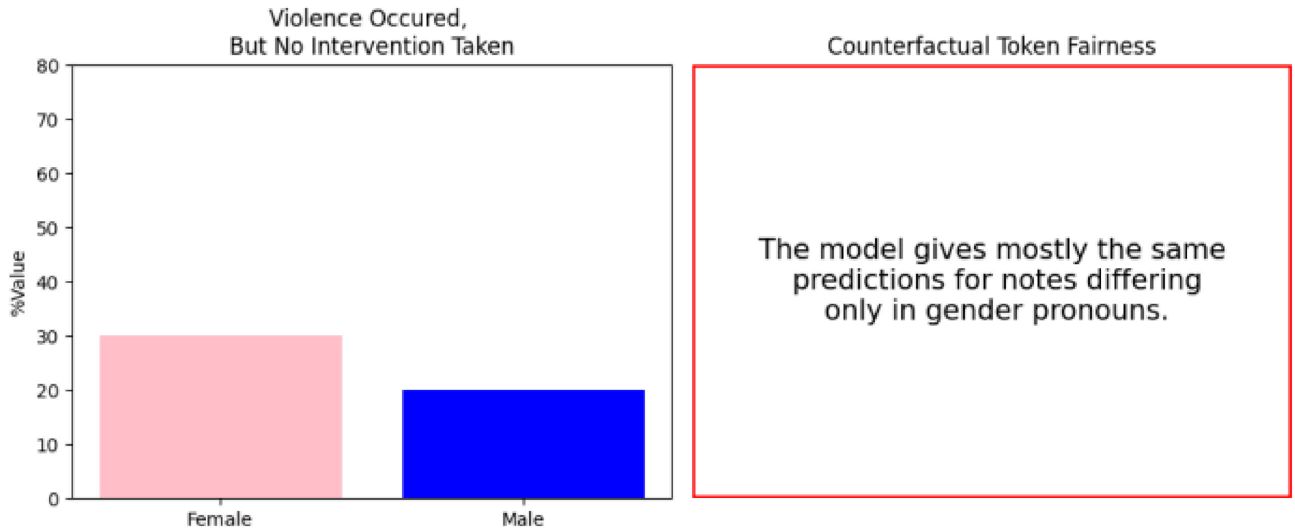
- Model 1
- Model 2
- Equally fair
- Equally unfair

* Which model is fairer; Model 1 or Model 2?

Model 1



Model 2



- ◯ Model 1
- ◯ Model 2
- ◯ Equally fair
- ◯ Equally unfair

Section 2: Demographics

Demographic Information

Almost done! This is the last section.
The following information will be used to describe your general demographics, without explicitly identifying you. You are free to skip any question.

Please specify the gender with which you most closely identify:

- Male
- Female
- Non-binary / third gender
- Prefer not to say

What is your age?

What is your country of residence?

Please add if you have any thoughts/feedback regarding survey.