4a.

PPV: Probability that a label is a true positive given the model predicted positively.

NPV: Probability that a label is a true negative given the model predicted negatively.

FPR: Probability that a label is positive given that the actual result was negative.

FNR: Probability that a label is negative given that the actual result was positive.

4b.

I would rather have a higher NPV because I would want the model to only deny loans in situations that are "true negatives" – this way I would be more confident that I wasn't being incorrectly denied a loan.

4c.

The authors recommend that in order to increase transparency and accountability, information should be provided about the demographic composition of datasets used to train/validate models and that the relative performance of algorithms on demographic subgroups should be tested, reported, and worked on if necessary. This relates to specific metrics in that accountability could look like comparing metrics (such as PPV or FPR) across different demographic groups and working to address differences.

4d.

The analysis of the authors is intersectional in that it separately considers combinations of traits across 2 different demographic axes (skin color and gender). The analysis finds that gender classification systems work best on males and light-skinned people and worst on females and darker-skinned people. The model works best with light-skinned males and worst with dark-skinned females.

4e.

Confounded in this context means that the results would be caused by differences in qualities of sensor readings rather than differences in skin color or gender. It is important to check this because the authors are claiming that there is bias *beyond* that caused by differences in clarity of photos taken based on skin color. As a result, they need to show that common ML algorithms still classify darker skin/women worse even when they attempt to choose only appropriately lit images (they do this by using high-quality parliamentary photos).