a)

Positive Predictive Value (PPV): PPV measures the proportion of positive predictions that are correct.

False Positive Rate (FPR): measures the proportion of actual negatives that are incorrectly classified as positives.

Negative Predictive Value (NPV): NPV measures the proportion of negative predictions that are correct.

False Negative Rate (FNR): FNR measures the proportion of actual positives that are incorrectly classified as negatives.

If I am applying for loan, I would rather have that decision made by a system with high FPR rather than a system with high FNR. With a system that has high False Positive Rate, meaning many individuals who should not be qualified as eligible applicant is falsely classify as eligible. Therefore, me as an applicant, would get a higher chance to be approved for the loan.

b)

I would rather to have a system with higher NPV instead of higher PPV. PPV focused on precision among approvals and measure the percentage of true positive among applicants. However, it may lead to denial to many qualified applicants to ensure confident in approvals. This is harmful to applicants. NPV, on the other hand, focused on precision among denial, and measure true positive, and minimizing false negative. It hence ensures that qualified applicant is rarely denied, which will directly benefit the applicants, ensuring fairness.

c)

Buolamwini and Gebru raise demands for accountability and transparency in the ML systems by calling for demographically balanced benchmark datasets that are inclusive, with intersectional auditing to evaluate performance in subgroups. Reporting of disaggregated metrics includes PPV, FPR, among others, to identify and address disparities across gender, race, and phenotypic subgroups. It means documentation of the composition of training data, model design, and performance metrics, whereas accountability would imply performance audits and mechanisms to address algorithmic failures. Metrics such as PPV and FPR are key in finding the biases to implement changes necessary for fairness in applications that are very sensitive, such as law enforcement and health. Their approach has emphasized how relevant the reduction of algorithmic performance gaps and fostering equity across different populations is.

d)

Buolamwini and Gebru's analysis is intersectional because it investigates performance disparities across overlapping demographic categories, in particular, gender (male, female) and skin type (lighter, darker), creating four subgroups: darker-skinned females, darker-skinned males, lighter-skinned females, and lighter-skinned males. Their findings indicate that commercial gender classifiers perform the worst for darker-skinned females with error rates up to 34.7%, whereas lighter-skinned males can have error rates as low as 0.0%, a significant gap of 34.4%. This analysis showcases the combined biases of groups at demographic intersections and emphasizes the role of using representative datasets and making targeted improvements to reduce disparities and ensure fairness in ML systems.

e)

In this context, "confounded" means that the observed disparities in gender classification accuracy could be influenced by external factors, such as differences in image quality (e.g., lighting, pose, or resolution), rather than biases in the machine learning systems themselves. Such a consideration is important to rule out confounding factors and validate findings, indeed, that the disparities are due to algorithmic and dataset biases, rather than artifacts of data collection. This supports the thesis of the paper that unrepresentative training data and flawed algorithmic design create intersectional biases, which lends credence to fairness-focused interventions through representative data and selective audits. Confirming the non-existence of confounding factors helps to ensure that the solutions proposed address the root causes of the problem.