In chapter 11, We learned about the key risks and mitigations of AI Ethics in security, Bias and False positives/Negatives being the risks and Subgroup evaluations and human approval being the mitigations. Biases appear when training data doesn't represent certain user groups which causes unfair treatment or unnoticed threats. Both false positives and negatives create risks, false positives will waste time analyzing and destroy trust while false negatives let potential threats pass by undetected. Subgroup evaluations test if the system performs equally across all different groups, granting opportunities to identify and correct hidden biases. Human approval provides a second analysis on alerts flagged by AI before critical actions are taken. These tie into the provided chapter as it focuses on the accountability, oversight and ensuring that AI remains a tool to assist and not do the job for you.

If I were to use AI in cybersecurity, I would be implementing controls to help maintain the effectiveness of AI without falling into over reliance on it. I would be making sure that there is a human in the loop review for escalated cases, periodic testing for bias and accuracy and a clear appeal process for users who are flagged. The appealing process will also include written explanations and timelines. Metrics will also be placed for false positives and negatives remaining below or at a certain percent and acknowledging appeals within 24 hours and resolved within a few days. Controls and metrics that would help maintain balance on efficiency, fairness and transparency.

A point that remains unclear throughout the reading is how much transparency is possible without exposing sensitive data. When an AI system flags a login attempt, how much of the reasoning should be shared with that user versus secured to protect how it's detected? Is an AI system's test enough to effectively test data while remaining confidential?

Relevancy:

- AI accountability matters as over reliance on automation can lead to not only unfair treatments of users but costly breaches.
- An example would simply be email servers flagging legitimate emails as suspicious content.