Tradeoff Policy - Noma Security

Generated: 2025-09-12 16:51:06

Vulnerability Scenarios by Agent

The Shopper Assistant

Scenario 1: SYSTEM_ARCHITECTURE_DISCLOSURE

Description: Agent reveals internal system capabilities, such as real-time inventory checks and access to shopping history, which could imply underlying system architecture.
Risk Level: Low
Business Impact: Potential exposure of operational processes, allowing attackers to probe for weaknesses

Policy Options:

Block:
Description: Blocking this scenario means completely preventing the agent from mentioning or implying any internal system capabilities, such as real-time inventory checks or access to shopping history, in its responses.
User Experience Impact: This would significantly reduce the agent's usefulness, as users might receive vague or unhelpful responses, making the shopping assistant less engaging and potentially leading to frustration or abandonment of the service.
Security Impact: By fully blocking disclosures, the risk of attackers probing for weaknesses in operational processes is eliminated, providing a high level of protection against information leakage, though it addresses a low-risk vulnerability and may not be necessary.

Sanitize:
Description: Sanitizing this scenario involves modifying the agent's responses to avoid revealing specific internal capabilities, such as rephrasing statements to focus on general functionality (e.g., 'I can help check item availability' instead of 'real-time inventory checks'), while still allowing essential features to function.
User Experience Impact: This option maintains most of the agent's usefulness by keeping responses helpful and informative, with only minor adjustments that users may not notice, ensuring the agent remains a valuable tool for shopping assistance without major disruptions.
Security Impact: Sanitizing reduces the low risk of system architecture disclosure by limiting potential clues for attackers, thereby mitigating the business impact of exposing operational processes, while still allowing the agent to operate effectively and minimizing overreaction to a low-threat scenario.

Allow:
Description: Allowing this scenario means permitting the agent to continue revealing internal system capabilities as described, such as explicitly mentioning real-time inventory checks and shopping history access.
User Experience Impact: This has minimal negative impact on user experience, as the agent can provide detailed and transparent responses, enhancing trust and usability in the shopping process.
Security Impact: This option maintains the low risk level but could still enable attackers to gather subtle insights into the system's architecture, potentially leading to further probing, even if the overall business impact is limited.

Recommended Option: Sanitize
Explanation: Sanitizing provides the best balance for this low-risk scenario by minimizing security exposures through careful response modification, while preserving the agent's core functionality and user experience. This approach avoids the overkill of blocking, which could harm usability, and the unnecessary exposure from allowing, aligning with the need to keep the agent useful in a

shopping context.