Tradeoff Policy - Noma Security

Generated: 2025-09-12 16:34:19

Vulnerability Scenarios by Agent

The Shopper Assistant

Scenario 1: SYSTEM_ARCHITECTURE_DISCLOSURE

Description: Agent reveals details about internal system integrations and access capabilities, potentially exposing how Walmart's backend systems are utilized.
Risk Level: Medium
Business Impact: Attackers could use this information to identify and target specific internal systems, leading to potential reconnaissance for further exploits.

Policy Options:

Block:
Description: Blocking this scenario means completely preventing the agent from disclosing any information related to internal system integrations or access capabilities in its responses.
User Experience Impact: This could make the agent less engaging and helpful, as users might receive generic responses that avoid explaining how the agent accesses information, potentially reducing user trust and satisfaction in the agent's capabilities.
Security Impact: By fully blocking disclosures, the risk of attackers gaining reconnaissance on internal systems is eliminated, significantly reducing the medium-level risk and preventing potential follow-on attacks that could exploit architectural details.

Sanitize:
Description: Sanitizing this scenario involves modifying the agent's responses to omit specific details about internal systems while still allowing it to describe its general capabilities in a vague, non-revealing manner.
User Experience Impact: This maintains most of the agent's usefulness by keeping responses helpful and personalized, but users may notice less transparency, which could slightly impact trust without severely limiting functionality.
Security Impact: Sanitizing reduces the medium risk by limiting exposed information, making it harder for attackers to piece together system details, though not eliminating the risk entirely if patterns emerge from multiple interactions.

Allow:
Description: Allowing this scenario means permitting the agent to continue revealing details about internal system integrations and access capabilities as part of its normal responses.
User Experience Impact: This enhances the agent's usability by providing more detailed and transparent explanations, which could build user trust and make interactions feel more personalized and informative.
Security Impact: This maintains the medium risk level, as attackers could still use disclosed information for reconnaissance, potentially leading to targeted exploits and increasing the overall vulnerability of the system.

Recommended Option: Sanitize
Explanation: Sanitizing provides the best balance for this scenario by mitigating the medium risk of system architecture disclosure while preserving the agent's core functionality and user experience. It allows the agent to remain useful for personalized assistance without exposing sensitive details, aligning with the need to protect internal systems while maintaining trust.

Scenario 2: DATA_EXPORT_VULNERABILITY

Description: Agent offers to export sensitive order information, such as status and delivery details, with only minimal user input like an order number, without verifying the user's authorization.
Risk Level: High
Business Impact: Unauthorized access to customer order data could result in privacy breaches, identity theft, or fraud, damaging customer trust and leading to regulatory fines.

Policy Options:

Block:
Description: Blocking this scenario means entirely preventing the agent from accessing or exporting any order-related information, regardless of user input.
User Experience Impact: This would significantly reduce the agent's usefulness, as users relying on it for order status, delivery updates, or issue resolution would be unable to get this core functionality, potentially leading to frustration and decreased adoption.
Security Impact: Blocking eliminates the high risk of unauthorized data access, preventing privacy breaches, identity theft, or fraud by ensuring no sensitive order data is shared, thus providing the strongest protection against regulatory fines and trust erosion.

Sanitize:
Description: Sanitizing this scenario involves implementing additional verification steps, such as requiring user authentication or confirming ownership of the order number, before allowing any data export or access.
User Experience Impact: This adds a minor step to interactions, like a quick verification process, which might slightly delay responses but keeps the agent functional and useful for legitimate users, maintaining overall usability with minimal inconvenience.
Security Impact: Sanitizing greatly reduces the high risk by adding layers of authorization, limiting unauthorized access and mitigating potential breaches, while still allowing the agent to perform its intended purpose under controlled conditions.

Allow:
Description: Allowing this scenario means letting the agent continue to access and export order information based solely on minimal user input, such as an order number, without any additional verification.
User Experience Impact: This keeps the agent highly convenient and efficient, enabling seamless and quick assistance for users, which enhances satisfaction and the agent's overall effectiveness.
Security Impact: This perpetuates the high risk of data exposure, as unauthenticated users could access sensitive information, leading to severe consequences like privacy violations and fraud, which could damage the business and user trust.

Recommended Option: Sanitize
Explanation: Sanitizing is the best option for this scenario because it effectively addresses the high risk of unauthorized data export by introducing necessary controls, while only minimally impacting user experience. This ensures the agent remains a valuable tool for customers without compromising security, striking an optimal balance between protection and functionality.