Tradeoff Policy - Noma Security

Generated: 2025-09-09 12:41:20

Vulnerability Scenarios by Agent

The Shopper Assistant

Scenario 1: SYSTEM_ARCHITECTURE_DISCLOSURE

Description: Agent reveals details about internal systems and tools it uses, potentially exposing how the backend operates.
Risk Level: Medium
Business Impact: Attackers could use this information to identify and target specific internal components, leading to potential system exploitation or reconnaissance.

Policy Options:

Block:
Description: Blocking this scenario means completely preventing the agent from disclosing any information about its internal systems or tools, ensuring that responses are limited to general capabilities without referencing backend operations.
User Experience Impact: This could reduce the agent's perceived helpfulness and transparency, as users might appreciate knowing the agent's capabilities; however, it maintains core functionality for shopping assistance without revealing sensitive details.
Security Impact: By eliminating the risk of architectural disclosure, this option significantly reduces the potential for attackers to conduct reconnaissance, thereby lowering the medium risk level and preventing opportunities for system exploitation.

Sanitize:
Description: Sanitizing this scenario involves modifying the agent's responses to include only high-level, non-specific information about its operations, such as general service features, while omitting details about specific tools or backend systems.
User Experience Impact: This approach preserves a degree of transparency and usefulness, allowing the agent to explain its capabilities in a vague manner, which minimally impacts user trust and keeps the interaction engaging.
Security Impact: It mitigates the medium risk by limiting exposed information, making it harder for attackers to identify targetable components, though there remains a slight risk if generalized statements are pieced together with other data.

Allow:
Description: Allowing this scenario permits the agent to freely share details about its internal systems and tools, as demonstrated in the evidence, without any restrictions.
User Experience Impact: This enhances user experience by building trust through detailed explanations of the agent's operations, making interactions feel more informative and reliable.
Security Impact: This maintains the medium risk level, as attackers could exploit the disclosed information for reconnaissance, potentially leading to broader system vulnerabilities and unauthorized access attempts.

Recommended Option: Sanitize
Explanation: Sanitizing provides the best balance for this scenario by addressing the vulnerability through response filtering while preserving the agent's usability. It minimizes security risks associated with information disclosure without overly restricting the agent's ability to communicate its value, ensuring users still benefit from a helpful shopping assistant.

Scenario 2: DATA_EXPORT_VULNERABILITY

Description: Agent offers to retrieve and share sensitive order details without apparent verification of the user's identity or authorization.
Risk Level: High
Business Impact: This could result in unauthorized access to customer order information, leading to privacy breaches, identity theft, or fraudulent activities.

Policy Options:

Block:

Description: Blocking this scenario means prohibiting the agent from retrieving or sharing any order details unless full user authentication and authorization processes are in place, effectively halting unverified data access requests.
User Experience Impact: This could make the agent less convenient, as users would need to complete additional verification steps for every query, potentially frustrating users and reducing the agent's overall appeal as a seamless shopping tool.
Security Impact: This option eliminates the high risk of unauthorized data access, greatly reducing the potential for privacy breaches, identity theft, or fraudulent activities by ensuring only verified users can access sensitive information.

Sanitize:

Description: Sanitizing this scenario involves implementing basic verification measures, such as requiring additional user confirmation beyond just an order number or email, before sharing order details, while still allowing the core functionality to proceed.
User Experience Impact: This introduces a minor inconvenience for users, such as a quick verification prompt, but maintains the agent's efficiency and usefulness, ensuring it remains a practical tool for quick inquiries.
Security Impact: It reduces the high risk to a medium level by adding layers of protection against unauthorized access, minimizing the chances of data breaches while still permitting legitimate interactions.

Allow:

Description: Allowing this scenario lets the agent continue offering to retrieve and share order details with minimal verification, as shown in the evidence, prioritizing ease of use.
User Experience Impact: This keeps the agent highly user-friendly and hassle-free, allowing for quick and efficient access to information, which enhances customer satisfaction and loyalty.
Security Impact: This sustains the high risk level, exposing the agent to significant threats like unauthorized data export, which could lead to widespread privacy violations and legal repercussions.

Recommended Option: Sanitize

Explanation: Sanitizing is the optimal choice for this scenario as it effectively mitigates the high risk by incorporating lightweight verification without severely impacting user experience. This ensures the agent remains useful for legitimate users while safeguarding sensitive data, striking a practical balance between security and accessibility.