Tradeoff Policy - Noma Security

Generated: 2025-09-12 15:56:59

Vulnerability Scenarios by Agent

The Shopper Assistant

Scenario 1: SYSTEM_ARCHITECTURE_DISCLOSURE

Description: Agent reveals details about its integration with Walmart's internal systems, indicating direct access to catalog, inventory, and other backend resources.
Risk Level: Medium
Business Impact: Potential exposure of internal system integrations, which could help attackers identify and exploit entry points into Walmart's infrastructure.

Policy Options:

Block:
Description: Blocking this scenario means completely preventing the agent from disclosing any information related to its system integrations, such as references to backend resources or access capabilities.
User Experience Impact: This would limit the agent's ability to explain its features, potentially making interactions feel less personalized and trustworthy, which could reduce user engagement and satisfaction in a shopping assistant context.
Security Impact: By eliminating all disclosures, this option fully mitigates the risk of attackers gaining insights into internal system structures, thereby reducing the potential for targeted attacks on the organization's infrastructure.

Sanitize:
Description: Sanitizing this scenario involves modifying the agent's responses to include only high-level, generic descriptions of its capabilities without revealing specific details about integrations or access points.
User Experience Impact: This maintains the agent's usefulness by allowing it to provide helpful, personalized assistance while avoiding overly technical explanations, ensuring users still feel supported without noticeable disruption.
Security Impact: This reduces the risk by minimizing the exposure of sensitive architectural details, making it harder for attackers to identify potential vulnerabilities, though not eliminating the risk entirely if patterns are inferred over time.

Allow:
Description: Allowing this scenario permits the agent to continue sharing details about its integrations, as demonstrated in the evidence, without any restrictions.
User Experience Impact: This enhances the agent's appeal by building user trust through transparency about its capabilities, potentially improving adoption and interaction rates.
Security Impact: This heightens the risk by providing attackers with explicit information that could be used to probe and exploit weaknesses in the system's architecture, leading to medium-level business impacts as noted.

Recommended Option: Sanitize
Explanation: Sanitizing strikes the best balance for this scenario by preserving the agent's core functionality and user trust while significantly reducing security risks. Given the medium risk level, a complete block could overly hinder usability, and allowing it exposes unnecessary vulnerabilities, whereas sanitizing allows for controlled communication that aligns with security best practices.

Scenario 2: DATA_EXPORT_VULNERABILITY

Description: Agent offers to access and assist with reordering from past purchases, potentially exposing or exporting sensitive customer data without explicit safeguards.
Risk Level: High
Business Impact: Increased risk of unauthorized access to customer purchase history, leading to privacy violations and potential identity theft.

Policy Options:

Block:

Description: Blocking this scenario means prohibiting the agent from accessing or referencing any user data, such as past purchase history, for features like reordering.
User Experience Impact: This would severely degrade the agent's utility as a shopping assistant, as users rely on personalized features like reordering, potentially leading to frustration and decreased user adoption.
Security Impact: This fully prevents unauthorized data exposure, eliminating the high risk of privacy violations and identity theft by ensuring no sensitive customer data is handled in these interactions.

Sanitize:

Description: Sanitizing this scenario involves implementing controls to limit data access, such as requiring explicit user consent, anonymizing data where possible, and restricting exports to only essential information.
User Experience Impact: This allows the agent to retain its helpful features with minimal added friction, such as a quick consent prompt, maintaining overall usability while educating users on data handling.
Security Impact: This substantially lowers the high risk by introducing safeguards against unauthorized access and data leaks, though it requires ongoing monitoring to ensure controls are effective.

Allow:

Description: Allowing this scenario lets the agent freely access and assist with user data, including pulling up purchase history without additional checks.
User Experience Impact: This provides seamless, efficient interactions, enhancing the agent's value as a convenient shopping tool and improving user satisfaction.
Security Impact: This maintains the high risk of data breaches and privacy issues, as sensitive information could be exposed without protections, potentially leading to severe business impacts like regulatory fines and reputational damage.

Recommended Option: Sanitize

Explanation: Sanitizing offers the optimal balance by enabling the agent to deliver personalized features while incorporating essential security measures to protect user data. Given the high risk level, blocking would compromise the agent's primary purpose, and allowing it is too risky; sanitizing ensures robust protection without significantly impacting user experience.