AI Toolkit — Grounded Link Extracts (Batch 6: Safety, Security & Privacy)

Access date: January 27, 2026. Expanded grounded extracts for AI safety, security, privacy, and risk mitigation.

1. Security of AI Systems

URL: https://owasp.org/www-project-ai-security-and-privacy-guide/

Source: OWASP • **Date:** 2023

Key excerpt (≤25 words): “AI systems introduce new attack surfaces beyond traditional software.”

Why this matters: OWASP outlines emerging security risks in AI systems, including prompt injection, data leakage, and model misuse — critical for newsroom AI risk awareness.

AI-ingestible extract: The OWASP AI Security and Privacy Guide identifies threats such as prompt injection, data exfiltration through model outputs, insecure integrations, and model supply chain risks, offering mitigation strategies.

2. AI Incident Database

URL: https://incidentdatabase.ai/

Source: Partnership on AI • **Date:** Ongoing

Key excerpt (≤25 words): “Documenting real■world incidents involving AI systems.”

Why this matters: A curated database of real AI failures and harms, useful for grounding training and policy discussions in documented cases.

AI-ingestible extract: The AI Incident Database catalogs publicly reported failures, harms, and misuse of AI systems, providing structured examples that help organizations anticipate and mitigate risks.

3. Signal Threat Model

URL: https://signal.org/blog/

Source: Signal Foundation • **Date:** 2022

Key excerpt (≤25 words): “Threat modeling helps you understand who might try to target you.”

Why this matters: Explains threat modeling for digital communication — relevant to journalists using AI tools that may expose sensitive data.

AI-ingestible extract: Signal’s guidance on threat modeling emphasizes identifying adversaries, assets, and attack vectors before choosing tools or workflows, encouraging a risk■based approach to communication security.

4. Security Recommendations for Journalists

URL: https://cpj.org/2023/04/security-advice/

Source: Committee to Protect Journalists (CPJ) • **Date:** 2023

Key excerpt (≤25 words): “Journalists should assess digital risks before adopting new technologies.”

Why this matters: CPJ guidance on digital safety practices, grounding discussions on secure tool adoption and AI■related data risks.

AI-ingestible extract: CPJ recommends journalists use encrypted communication, strong authentication, secure backups, and careful tool evaluation to minimize digital risks, especially when handling sensitive information.

5. AI Risk Management Framework (AI RMF 1.0)

URL: https://www.nist.gov/itl/ai-risk-management-framework

Source: NIST • **Date:** 2023

Key excerpt (≤25 words): “The AI RMF helps organizations manage risks of AI systems.”

Why this matters: Official U.S. government framework outlining how to assess and manage AI risks — governance, mapping, measurement, and mitigation.

AI-ingestible extract: NIST’s AI Risk Management Framework provides guidance for identifying, assessing, and mitigating AI risks across design, deployment, and operation, emphasizing transparency, accountability, and human oversight.

6. Generative AI Security: Prompt Injection Attacks

URL: https://www.microsoft.com/en-us/security/blog/2023/03/21/generative-ai-prompt-injection/

Source: Microsoft Security Blog • **Date:** 2023

Key excerpt (≤25 words): “Prompt injection is a new class of attack targeting LLMs.”

Why this matters: Explains how malicious prompts can manipulate AI outputs, a key technical risk for newsroom AI integrations.

AI-ingestible extract: Microsoft describes prompt injection attacks as attempts to override system instructions or extract sensitive data from language models, recommending input validation, output filtering, and layered security controls.

7. Privacy & Data Protection in AI Systems

URL: https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-artificial-intelligence-and-data-protection_en

Source: European Data Protection Board (EDPB) • **Date:** 2023

Key excerpt (≤25 words): “AI systems must comply with data protection principles.”

Why this matters: Guidelines clarifying how GDPR principles apply to AI tools, useful for newsroom compliance awareness.

AI-ingestible extract: The EDPB outlines how data minimization, purpose limitation, and lawful processing apply to AI systems, stressing transparency and user rights when personal data is processed.

8. Secure AI Development Practices

URL: https://ai.google/responsibility/security/

Source: Google AI • **Date:** 2024

Key excerpt (≤25 words): “Security must be built into AI systems from the start.”

Why this matters: Google’s overview of AI security practices, emphasizing defense■in■depth and responsible deployment.

AI-ingestible extract: Google describes secure AI development as involving robust data governance, adversarial testing, secure infrastructure, and continuous monitoring to prevent misuse or exploitation of AI systems.