🚨 2025 IR Update Checklist 🚨
AI Incident Response

☀️ Sandy Dunn

1. 🛠️ Preparation

- ☐ Add AI specific incident types to your IR policy (misinformation prompt injection, hallucinations, deep fakes, etc.).
- ☐ Maintain an updated AI asset inventory (models, datasets, pipelines, agents).
- ☐ Define clear roles for AI incident handling (AI Security Lead, MLOps, Privacy Officer).
- ☐ Develop AI-specific playbooks, severity levels, and escalation paths.
- ☐ Integrate insights from AI red teaming and threat modeling.

2. 🔎 Detection & Analysis

- ☐ Expand telemetry: capture prompt logs, inference results, agent activities.
- ☐ Define triage rules for AI-specific incidents like jailbreaks, output toxicity, agent misbehavior.
- ☐ Use AI-aware severity classifications and behavioral baselines.
- ☐ Monitor for deep fakes, model drift, adversarial prompts, and data leakage.

3. 🔒 Containment

- ☐ Be able to pause or disable AI systems, revert model versions, or quarantine agents.
- ☐ Implement prompt hardening or policy restrictions dynamically.
- ☐ Respond to misinformation using Break Out Scale Guidelines
- ☐ Stop spread of generative content misuse (e.g., fake images, cloned voices).
- ☐ Disable access to compromised APIs, vector stores, or toolchains.

4.🦟 Eradication

- ☐ Identify the root cause: training data issues, prompt flaws, supply chain compromise.
- ☐ Retrain or rollback models as needed; replace vulnerable prompts or pipelines.
- ☐ Deploy canary prompts to validate behavior.
- ☐ Conduct red team testing postfix.

5. 🌱 Recovery

- ☐ Restore aligned, safe model and agent operation.
- ☐ Re-enable logs and ensure output safety.
- ☐ Communicate clearly if decisions were impacted by bad AI output.
- ☐ Debrief engineers and security teams.

6. 📈 Post-Incident Review

- ☐ Run an AI aware root cause analysis (RCA).
- ☐ Update threat models and prompt guidelines.
- ☐ Capture incident metrics and lessons learned.
- ☐ Train teams on findings; update documentation and governance models.

🚀Additional Best Practices

- ☐ Simulate AI incidents via red team/tabletop exercises.
- ☐ Cross-train SOC/IR teams on LLM-specific threats.
- ☐ Track KPIs such as MTTR for AI incidents and percentage caused by third-party components.