AI Incident Response Plan

Maintained by: City of San José's Information Technology Department Digital Privacy Office
digitalprivacy@sanjoseca.gov

Last updated: 7/1/2024

Introduction

The AI Incident Response Plan (IRP) serves as the first line of defense for the City of San José in case of an AI incident. This IRP has been created based on the NIST AI Risk Management Framework and the Special Publication 800-61 Computer Security Incident Handling Guide.

Incident response occurs in sequential phases, each one building upon the next. The following phases provide a foundation for an Incident Response (IR) Team to respond to and recover from an AI incident:

1. Preparation
2. Detection & Analysis
3. Containment
4. Eradication & Recovery
5. Post-incident Activity

Purpose

The purpose of this document is to prepare for and gain a fundamental understanding of the processes, responsibilities, and actions required to mitigate an AI incident. It is critical to identify and resolve incidents quickly before they escalate into a major incident with the potential to cause harm or damage to people, data, or the City of San José.

Scope

This IRP applies to all AI systems implemented by City of San José (City) staff, contractors, and any entity operating on behalf of the City. The AI IRP addresses continuity and recovery procedures to appropriately mitigate AI incidents.

Approach

The key points in development of the AI IRP include:

- **Evaluate**: Evaluate risk levels and determine the appropriate response for an AI incident, which may include obtaining senior management support.
- **Plan**: Keep the plan simple. A well-organized, systematic, and up-to-date AI IRP that is readily available will help teams get through most situations.

- **Communicate**: Communicate regularly on the incident status. Provide the relevant facts as they are available, disseminate them quickly, follow up regularly, keep relevant parties informed and resolve incorrect information.
- **Review**: Review the AI IRP at least bi-annually to ensure the documented procedures are still appropriate and that the team is equipped to respond accordingly. Ensure an after-action plan is developed and communicated, with assigned improvement opportunities.
- **Test**: The AI IRP should be tested with tabletop exercises at least annually.
- **Be flexible**: The AI IRP should exhibit flexibility to meet a wide variety of situations, including modular team membership and access to the appropriate resources. External partners such as law enforcement should be involved as needed.

Definitions

The definitions in this section are based on the Responsible AI Collaborative's AI Incident Database.[1]

AI Incident: An AI incident is an alleged harm or near harm event to people, property, reputation, technical integrity of the environment where an AI system is implicated, or the City. Examples of AI incidents include an AI system providing false information, engaging in copyright infringement, generating harmful bias, being misuses, exposing confidential or sensitive information, and posing a liability risk.

Harm: Allegations of an AI incident must be assessed by the AI IR Team, erring on the side of finding harm when there is a plausible argument that harm has occurred. In certain cases, this requires consultation with attorneys and executives. Types of harm include, but are not limited to:

- Harm to physical health/safety
- Psychological harm
- Financial harm
- Harm to physical property
- Harm to intangible property (e.g., IP theft, reputation damages)
- Harm to social or political systems (e.g., election interference, loss of trust in authorities)
- Harm to civil liberties (e.g., unjustified imprisonment or other punishment, censorship)

Harms do not have to be severe to meet this definition; an incident resulting in a minor, easily remedied expense or inconvenience would still be considered a harm for our purposes.

[1] https://incidentdatabase.ai/

Near harm: A near harm AI incident occurs when an AI system plays an important role in a chain of events that easily could have caused harm, but some external factor kept the harm from occurring. This external factor should be independent of the AI system and should not have been put in place specifically to prevent the harm in question.

Roles & Responsibilities

AI Incident Response Team

An AI Incident Response Team is established prior to any incidents and provides a quick, effective, and orderly response to AI incidents. The AI IR Team's mission is to address, handle, and resolve allegations of any harm caused by an AI incident.

The AI IR Team is authorized to take appropriate steps deemed necessary to contain, mitigate, or resolve an AI incident. The Team should have a charter, assigned roles and responsibilities, and a vetted incident response checklist. The Team is responsible for investigating any malfunctioning components of City of San José AI systems in a timely, cost-effective manner and reporting findings to management and the appropriate authorities as necessary. The City Privacy Officer (CPO) shall coordinate these investigations.

Contact Information

The AI IR Team shall include the following members:

- City Privacy Officer: digitalprivacy@sanjoseca.gov
- Cybersecurity team: cybersecurityteam@sanjoseca.gov
- Cybersecurity office: CybersecurityOffice@sanjoseca.gov

Other Teams as Needed

Other teams and departments shall contact, respond to, and support the CPO with any information relating to a suspected AI incident.

Activation Criteria

The CPO will determine whether the reported incident is serious enough to warrant an AI incident response. The following events responsible for causing alleged harm or near harm require reporting to the CPO and may trigger an AI incident response:

- Declining accuracy rates; includes false positives, false negatives, true positives, and true negatives;
- Findings or suspicions of algorithmic bias;
- Failed or broken human oversight protections;
- Failure of an AI system to provide its expected outcome (i.e., system has stopped working);
- Suspected or demonstrated harm to human livelihood; or

- Discovery of City of San José confidential data used to train any public AI model, especially in large language models or other generative AI services.

Some of these events may be connected to a broader incident that also involves cybersecurity, legal, or privacy teams. When unclear, contact both the CPO, the Chief Information Security Officer (CISO), and the Office of the City Attorney.

Response Level

Once an incident has been identified, the next step is to determine the level of response required based on its assessed risk and potential for harm. The City of San José should establish and implement a vetted risk matrix to assist with alleged AI harm. Determining the level of response is critical, as it dictates the involvement of different layers of the organization, including elected officials and internal communications teams as appropriate, as well as other external entities such as insurers, law enforcement, news media, and others. Each response level includes the following:

- **Scope:** An overview defining what each corresponding risk level entails.
- **Management Team:** The departments, individuals, and stakeholders responsible for managing the resolution of the incident.
- **Initial Bridge Convened:** The initial communication channel(s) used to connect involved parties.
- **Triggers for Escalation to the Next Level:** The activation requirements to raise the risk level of an incident.

For all levels of risk, City of San José staff, contractors, any entity operating on behalf of the City of San José, residents, and external organizations may report potential AI incidents via email. The AI IR Team may use Zoom, Microsoft Teams, and email to resolve reports.

Level 1: Low Risk
The scope of a low-risk AI incident is relatively small, usually involving minor systems errors with limited harm or near-harm to people. Examples include reporting an AI system that has a high accuracy rate but has started to output incorrect recommendations sporadically or an AI system that no longer provides a clear value to the City of San José. A low-risk incident does not impact any areas of human livelihood (e.g., civil rights, healthcare, employment, recreation, etc.).

- **Incident Management Team:** AI IR Team and AI system owner (if applicable)
- **Notification to:** Chief Information Officer (CIO), CPO, CISO, Department AI representative (if applicable)
- **Initial Bridge Convened:** These are routine or easily solvable incidents that do not require a bridge.
- **Triggers for Escalation to the Next Level:** If the impacted system impacts human livelihood, becomes increasingly inaccurate or indicates a need for sunsetting.

Level 2: Medium Risk
A medium-risk AI incident typically involves moderate levels of harm or near harm, such as those caused by periodic gaps in human oversight or frequent system errors. Examples include an AI system that is consistently inaccurate some of the time or sporadically does not allow a human to override its output. A medium-risk incident usually impacts non-critical areas of livelihood (e.g., recreation).

- **Incident Management Team:** AI IR Team and CIO
- **Notification to:** CISO, department AI representative (if applicable), department head
- **Initial Bridge Convened:** The CPO sends a meeting invite to all parties involved.
- **Triggers for Escalation to the Next Level:** If the impacted system fails to work or impacts critical areas of human livelihood.

Level 3: High Risk
A high-risk AI incident typically involves high levels of harm or near harm, such as those caused by failures in human oversight or major system errors. Examples include privacy violations, algorithmic bias and recommendations that cause harm. A high-risk incident usually includes clear evidence of suspected or demonstrated harm to critical areas of human livelihood like an individual's civil rights, healthcare, or employment.

- **Incident Management Team:** City Manager's Office, CIO, CISO, AI IR Team, OER, HR, Public Relations, Department Head, Elected Officials, CAO, CBO
- **Initial Bridge Convened:** The CPO, or equivalent sends a meeting invite to all parties involved.

External Communications

Organizations may have a need to communicate with external parties throughout the course of an AI incident. External parties may include media, constituents, external subject matter experts, technology vendors, other incident response teams, and law enforcement.

Key activities to establishing an effective external communication plan include:

- Plan incident communications procedures with the City of San José's public affairs, legal, and executive teams prior to an incident occurring and understand the role of each external party.
- Consult with legal and public affairs before initiating external communications to comply with any established contractual agreements or communication protocols.
- Consider communication cadence throughout the AI incident lifecycle, rather than waiting until the incident is resolved.
- Establish the organization's communications preference: ad-hoc (e.g. human generated email), semi-automated or automated.

- Exercise caution when sharing information externally. The AI IR Team will use its discretion to publicly release information in accordance with existing City of San José policy.
- Determine the relevant technical information to share externally. General characteristics of the incident may be safe to publicly release, whereas information about exploited vulnerabilities may pose considerable privacy and cybersecurity risks to the City of San José.

Phased Approach

The AI IRP follows a phased approach in accordance with the incident response cycle from the NIST Special Publication 800-61 Computer Security Incident Handling Guide. Each phase is discussed in more detail in the following sections.

1. Preparation
2. Detection & Analysis
3. Containment
4. Eradication & Recovery
5. Post-incident Activity

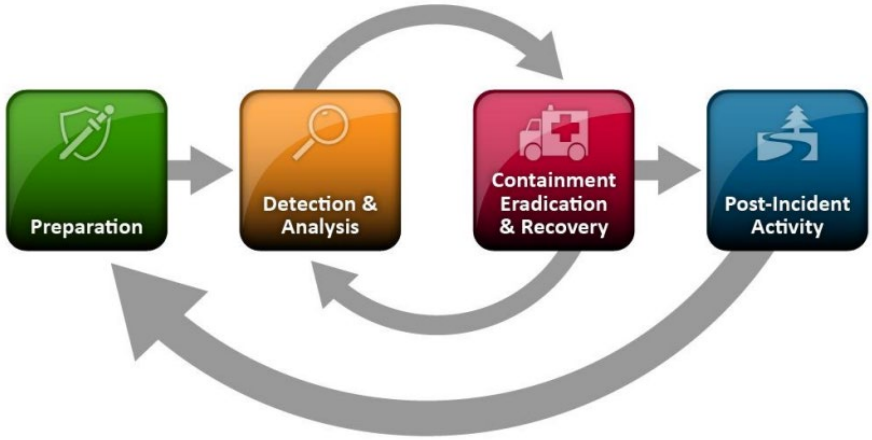

Figure 1: Incident Response Life Cycle[2]

1. Preparation

The preparation phase takes place before an incident occurs. Putting more effort into preparation can help reduce the duration and difficulty of recovering from an incident. Planning to respond to incidents involves all stakeholders with responsibilities to maintain the information system.

AI IR team members should avoid spending valuable time locating and assembling needed tools or supplies when an incident occurs by having them ready at all times. Supplies to

[2] https://nvlpubs.nist.gov/nistpubs/specialpublications/nist.sp.800-61r2.pdf

assist the team in the event of an incident are sometimes referred to as a "jump bag", and should include (at a minimum):

- An empty notebook and writing instruments (in the event that technological systems are compromised, thorough documentation should be taken throughout an incident).
- Contact information sheet.
- At least one computing device (e.g. laptop) for any technical support needed to address an incident. This laptop should be scrubbed and all software reinstalled before it is used for another incident.
- At least one other computing device (e.g. laptop) for writing reports, reading email, and performing other duties unrelated to the hands-on incident analysis.

Having contact information available and current is essential. It should also be in hard copy form in case electronic systems are affected during the incident. Contact information should include:

- All members of IR team
- All contacts listed in notification
- Personnel that might assist in handling an incident

In the event that critical systems relying on the implementation of AI become compromised, management team members are responsible for ensuring there is an alternative solution or backup system while the incident is being resolved.

Key activities during preparation include:

- Ensure incident response efforts are unified and coordinated.
- Implement, document, and audit maintenance procedures for each AI system regularly to reduce the likelihood, severity, and duration of AI incidents.
- Assess risk and disaster resilience to reduce the duration of AI incidents.
- Follow the City of San José's disaster recovery plan should an AI incident require it.
- Identify the potential harms of AI systems; incorporate this into the preparation processes.
- Establish a Critical Infrastructure Protection Plan and Business Continuity Plan to ensure critical infrastructure remains operational and business operations are sustained during incident response activities.
- Train responders to handle potential incidents should they arise. Assemble required tools and ensure responders know how to use them.
- Conduct tabletop exercises at least annually to keep the AI IR Team and IR Plan optimized. Exercises involving simulated incidents can be beneficial in preparing the IR Team for incident handling as well as refining AI IR Plan.
- Prepare and approve communications (internal, website, elected officials, and media).

Enhancing general user awareness across the City of San José is likely to reduce the occurrence of AI incidents. User awareness may be enhanced by the following activities:

- Create an AI awareness training program. The AI IR Team may incorporate it into the City of San José's existing Cybersecurity and Privacy awareness training programs. These trainings should occur at least annually.
- Ensure all users are aware of all AI and IT policies, including where current policy documents may be found.
- Share information about past AI incidents with end-users so that future occurrences may be prevented.

2. Detection & Analysis

Awareness that an AI incident has occurred can originate from different sources such as vendors, City of San José staff, or even residents. **An AI incident shall be defined by satisfying at least one of the conditions outlined in the previous section, "Activation Criteria".**

The primary outcome of the identification phase is to determine whether an event constitutes an AI incident, activate the IR plan, and begin first level notification.

When the CPO determines that an adverse risk exists, the CPO shall declare that an incident has occurred and assemble the AI IR Team to implement the IR Plan. To save all key records and begin detailed documentation, all relevant stakeholders should be involved as early as possible. This should include the City of San José's legal team to guide the investigation and determine the legal privilege and liability involved.

Key activities during this phase include:

- Identify, assess, and prioritize risks.
- Ensure the capacity for timely communications in support of situational awareness and operations during an AI incident.
- Provide all decision makers with decision-relevant information regarding the nature and extent of the AI incident, any cascading effects, and the status of the response.

As soon as the AI IR Team suspects that an incident has occurred, it is important to immediately start gathering all the relevant data for the relevant staff with provide access to this information to authorized City staff. Each incident is carefully documented and all related artifacts are kept for future reference.

3. Containment

After an AI incident has been identified, it is imperative to limit damage from the incident and isolate the affected components of the AI systems to prevent further damage. Depending on the type of incident, the overall strategy for containment may vary. Factors that help determine the containment strategy include:

- Potential damage to City of San José services
- Need for evidence preservation
- Service availability
- Time and resources needed to implement the strategy
- Effectiveness of the strategy (e.g., partially contains the incident, fully contains the incident)
- Duration of the solution (e.g., emergency workaround to be removed in four hours, temporary workaround to be removed in two weeks, permanent solution)

Evidence handling must be considered when resolving an incident. This will help in legal proceedings. In such cases, it is important to clearly document how all evidence has been preserved. In addition, evidence must be accounted at all times; whenever evidence is transferred from person to person, chain of custody documentation should detail the transaction and include each party's signature. A detailed log should be kept for all evidence including, but not limited to, the following:

- Identifying information (location, serial number, model number, hostname, media access control (MAC) address, and IP address of a computer)
- Name, title, phone number of each individual who collected or handled the evidence during the investigation
- Time and date of each occurrence of evidence handling
- Locations where evidence was stored

4. Eradication & Recovery

Identify the AI incident's root cause through a detailed forensic analysis and pause the usage of affected systems. Once the threat has been contained and the initial cause has been determined, second level notification can take place.

Key activities during eradication include:

- Prioritize the stabilization of critical infrastructure functions.
- Document everything, including the identified root cause of the incident.
- Implement any necessary safeguards to prevent the same kind of AI incident from occurring again.

Key activities to restore affected systems according to the configurations outlined in the AI Factsheet include:

- Verify affected systems meet their original objective.
- Test, validate, and approve affected systems before bringing them back into production in accordance with the City of San José's change management policies.
- Allow affected systems back into use and ensure no problem remains.

5. Post-incident Activity

Document the incident and analyze how it happened so staff can learn from it and improve future response efforts. Once documentation is complete, communicate the lessons learned to relevant stakeholders involved with the incident.

Items to review include:

- Consider whether an additional policy could have prevented the incident.
- Consider whether a procedure or policy was not followed, and then consider what could be changed to ensure that the procedure or policy is followed in the future.
- Was the incident response appropriate? How could it be improved?
- Was every appropriate party informed in a timely manner?
- Were the incident-response procedures detailed and did they cover the entire situation? How can they be improved?
- Have changes been made to prevent a new and similar incident?
- Should any policies or procedures be updated?
- What lessons have been learned from this experience?
- Create a follow-up action plan with due dates and assignments.
- What additional tools or resources are needed to detect, analyze, and mitigate future incidents?