The County of San Diego Artificial Intelligence Incident Response Plan

Summary:

The AI Incident Response Plan for the County of San Diego outlines a comprehensive approach to addressing incidents related to AI bias, security breaches, and unintended policy implications. The plan includes preparation through team formation and training, identification and documentation of incidents, immediate containment measures, thorough investigation to determine root causes, implementation of corrective actions, system recovery, and continuous improvement. This structured response ensures the protection of residents' rights and the integrity of AI systems used by the County.

1. Preparation

- **Incident Response Team**: The incident response team is comprised of members from IT, legal, compliance, human resources, and communications. This team will be responsible for managing AI-related incidents and will be led by the Chief Information Security Officer. At minimum this team shall include the following positions: the Chief Information Officer, the Chief Information Security Officer, and the Compliance Officer.
- **Roles and Responsibilities**: Roles and responsibilities are defined for each team member to ensure efficient coordination during an incident.
- **Training and Awareness**: Incident response training sessions are conducted to ensure all stakeholders are aware of AI risks, response protocols, and their roles in the incident response process.
- **Monitoring Systems**: Continuous monitoring and alerting systems are used to detect potential incidents early. This includes anomaly detection tools and regular audits of AI systems, specifically data loss prevention (DLP) systems.

2. Identification

- **Incident Detection**: Automated tools and manual reviews are in place to identify incidents related to AI bias, security breaches, and unintended policy implications. This includes monitoring system logs, user reports, and performance metrics.
- **Severity Assessment**: Each incident is evaluated based on its impact on residents, systems, and operations. Incidents are categorized into levels (e.g., low, medium, high) to prioritize response efforts similar to any other IT incident.

- **Documentation**: Detailed records of the incident are maintained including time, nature, affected systems, and initial assessment. This documentation is used for tracking the incident and informing subsequent steps.

3. Containment

- **Immediate Actions**: Using the County's IT Incident Response Framework, actions are taken immediately to contain an incident and prevent further damage. These actions may include isolating affected systems, suspending AI operations temporarily, or rolling back to previous versions of algorithms.
- **Communication**: County leadership and team members are quickly informed about incidents and containment measures. Additionally, notifications are sent to residents and employees of San Diego County in the case an incident has impact to the public. Transparency is key to maintaining trust.

4. Investigation

- **Root Cause Analysis**: A thorough investigation is conducted to determine the root cause of the incident. This involves analyzing system logs, data inputs, algorithm behavior, and any external factors.
- **Bias Detection**: For AI bias incidents, data and algorithms are analyzed to identify sources of bias and discrimination. This may involve reviewing training data, model parameters, and decision-making processes.
- **Security Breach Analysis**: For security breaches, assessments are conducted regarding the extent of unauthorized access and data compromise. The assessment includes examining the breach timing, affected data, and compromised vulnerabilities.
- **Policy Implications Review**: Unintended policy implications and their impact on residents and operations are evaluated by assessing whether AI decisions align with County policies and identifying any discrepancies.

5. Mitigation

- **Corrective Actions**: Corrective measures are implemented to address the root cause and prevent recurrence. This may include updating algorithms, enhancing security protocols, revising policies, or retraining models with unbiased data.

- **Communication**: Updates to stakeholders are provided about the mitigation steps taken, which includes affected residents to inform them how their concerns are being addressed.

6. Recovery

- **System Restoration**: Affected systems are restored, and functionality is validated as operating correctly and securely before resuming normal operations.
- **Data Integrity**: The post-incident integrity and security of data is ensured through automated and manual verifications to ensure no data has been tampered with or lost during the incident.
- **Long-term Fixes**: Long-term solutions are considered and implemented to strengthen AI systems against future incidents. This may involve adopting new technologies, improving monitoring systems, or revising incident response protocols.

7. Review and Improvement

- **Post-Incident Review**: Reviews of the incident response process are conducted to identify areas for improvement including gathering feedback from all involved parties to understand what worked well and what could be improved.
- **Lessons Learned**: Lessons learned are documented from the incident and the incident response plan is updated accordingly. These insights are shared with relevant stakeholders to enhance overall preparedness.
- **Continuous Improvement**: The AI incident response plan is regularly updated and reviewed to adapt to new threats, technologies, and regulations to ensure the plan remains effective and relevant in a rapidly evolving AI landscape.