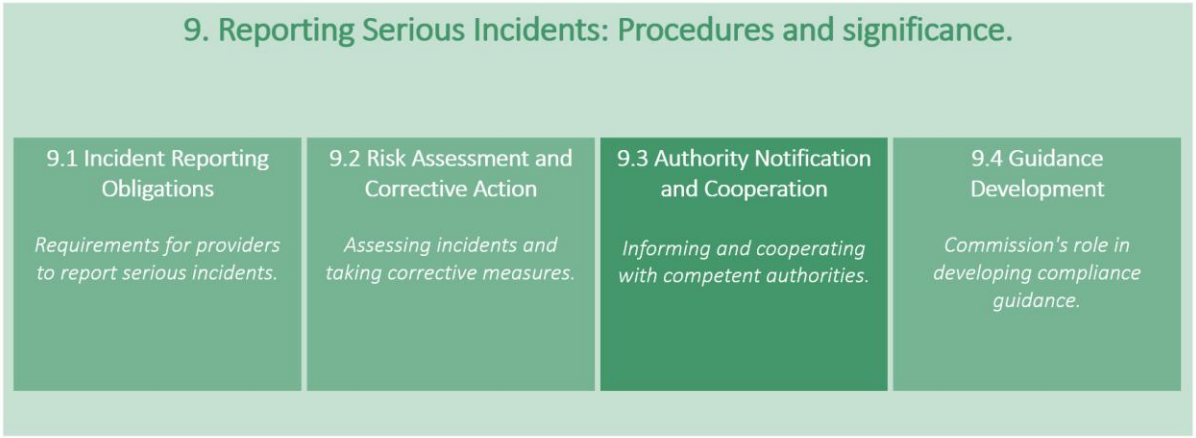
Strengthening AI Safety: The Pathway for Reporting Serious Incidents and Collaborating with Authorities

Co-authored with Uthman Ali, *Global AI Ethics & Safety Leader*

10 February 2025



Introduction

This is the third article in our series about reporting serious incidents under the EU AI Act. The first article in this series covered Incident Reporting Obligations with Dr. Benedikt Kohn, swiftly followed by Risk Assessment and Corrective Action with Matt Hervey.

To begin, we look at the basics: What is the EU AI Act? The EU AI Act is a landmark regulation designed to govern AI systems, ensuring safety, compliance, and innovation across the European Union.

In today's rapidly evolving landscape of artificial intelligence (AI), the accurate reporting of serious incidents stands as a cornerstone of ensuring both public safety and regulatory compliance. Under the EU AI Act, this reporting mechanism forms a crucial component in safeguarding individuals and organizations from potential harm arising from AI system malfunctions or misuse.

Central to this process is the collaborative effort between AI system providers and competent authorities. By fostering open communication and cooperation, both parties can work together to swiftly address incidents, assess risks, and implement necessary corrective actions. This collaborative approach not only enhances the effectiveness of incident response but also promotes transparency and accountability within the AI industry.

Moreover, the EU AI Act underscores the importance of proactive engagement between providers and authorities, highlighting the shared responsibility in upholding safety standards and regulatory requirements. By recognizing the critical role of cooperation, stakeholders can collectively contribute to strengthening AI safety measures, ultimately fostering trust and confidence in AI technologies.

In this article, we explore the essential pathway for reporting serious incidents under the EU AI Act and examine the pivotal role of collaboration with authorities in enhancing AI safety standards. Through a clear understanding of these processes and the importance of cooperation, stakeholders can navigate the complex landscape of AI regulation with confidence and responsibility.

The EU AI Act: A Primer on Serious Incident Reporting

The EU AI Act establishes clear guidelines for reporting serious incidents involving high-risk AI systems. Providers of such systems are obligated to promptly report any serious incidents to the market surveillance authorities of the Member States where the incident occurred.

This reporting requirement serves as a fundamental aspect of ensuring AI safety and regulatory compliance within the European Union. By promptly notifying authorities of serious incidents, providers contribute to swift and effective response measures, mitigating potential risks and safeguarding public safety.

Under the EU AI Act, the focus lies on proactive engagement and transparent communication between providers and competent authorities. This collaborative approach enables stakeholders to address incidents comprehensively, assess associated risks, and implement necessary corrective actions promptly.

In adhering to the reporting obligations outlined in the EU AI Act, providers demonstrate their commitment to upholding safety standards and regulatory requirements. Additionally, by fostering cooperation with authorities, providers can enhance AI safety measures and contribute to building trust and confidence in AI technologies.

In essence, the EU AI Act's provisions for reporting serious incidents establish a crucial pathway for strengthening AI safety. Through timely and transparent reporting, providers play a pivotal role in promoting accountability and ensuring the responsible deployment of AI systems across the European Union.

Real-World Example: The Google Gemini Incident

The Google Gemini incident—in which the AI tool produced offensive and historically inaccurate images—serves as a potential example of the type of malfunction that would necessitate reporting under the EU AI Act. Accordingly, generating offensive and culturally insensitive depictions of historical figures, the AI system breached public trust and caused reputational damage to its provider. These outputs could qualify as "serious incidents" under the EU AI Act, as they undermine societal values and have the potential to cause widespread harm. Given that the system's outputs directly impacted sensitive cultural and historical narratives, it highlights a failure in the system's safety and reliability. Such a high-profile failure would require immediate reporting to the competent authorities for investigation and corrective action.

The failure of Gemini's internal controls, such as safeguards against biased or inaccurate content generation, raises significant questions about the adequacy of the model's design and testing procedures. Under the EU AI Act, this lack of robustness may demand scrutiny and reporting. Google Gemini integrates both large language models (LLMs) and image-generation systems, representing a sophisticated example of generative AI technology. Such systems are increasingly – but no means certain – classified as high-risk under the EU AI Act due to their potential to influence public perception and decision-making. Consequently, the failure of Gemini underscores the importance of rigorous testing, transparency, and accountability in the deployment of generative AI systems. In retrospect, this could be breach subject to Article 73, yet more clarity from law enforcement is needed at this point.

Drawing Parallels with GDPR

Both the EU AI Act and GDPR emphasize accountability as a core point of compliance. Organizations are required to demonstrate proactive efforts to mitigate risks and report incidents promptly, ensuring transparency and public trust. Just as GDPR mandates reporting breaches within 72 hours, the EU AI Act also imposes strict deadlines for notifying authorities of serious incidents. These timelines are designed to facilitate swift responses and prevent further harm. Failure to adhere to either regulation can result in significant fines, reputational damage, and potential legal action. This shared focus on enforcement highlights the importance of regulatory compliance in both contexts.

However, there are notable distinctions between the two frameworks. GDPR breaches typically involve the unauthorized access, loss, or misuse of personal data, whereas AI incidents often pertain to malfunctions, safety risks, or ethical concerns arising from the operation of high-risk AI systems. AI incident reporting also involves a deeper analysis of causal links between system malfunctions and their consequences, often requiring advanced technical expertise. GDPR, while complex, is more focused on data protection and privacy. Additionally, while GDPR breaches directly affect individuals whose data has been compromised, AI incidents may have broader societal implications, influencing public discourse, cultural narratives, or even democratic processes.

Notification Procedures and Timelines

The EU AI Act establishes clear procedures and timelines for reporting serious incidents involving high-risk AI systems. Providers are required to promptly notify the market surveillance authorities of the Member States where the incident occurred.

Once a causal link between the AI system and the incident is established, immediate reporting is mandatory. The Act stipulates varying deadlines for reporting, depending on the severity of the incident.

For general incidents, providers must report within a maximum of 15 days. However, for more serious incidents, such as those resulting in death or widespread infringements, the reporting deadline is shorter, typically within two days.

These notification procedures ensure swift and effective response measures, enabling authorities to assess risks and take appropriate actions promptly. By adhering to these timelines, providers demonstrate their commitment to upholding safety standards and regulatory compliance within the EU AI landscape.

In summary, the EU AI Act's notification procedures and timelines establish a structured pathway for reporting serious incidents. Compliance with these requirements is essential for strengthening AI safety measures and fostering trust in AI technologies across the European Union.

Cooperation with Competent Authorities

Providers are mandated by the EU AI Act to collaborate closely with competent authorities during the investigation of serious incidents involving high-risk AI systems. This cooperation is crucial for ensuring thorough and effective investigations that uphold public safety and regulatory compliance.

Central to this obligation is the prohibition against altering the AI system in a manner that could influence the evaluation of the incident's causes before informing the competent authorities. This ensures the integrity of the investigation process and prevents any interference that may compromise the accuracy of findings.

As a result of adhering to these requirements, providers demonstrate their commitment to transparency, accountability, and proactive engagement in addressing serious incidents. This collaborative approach fosters trust between providers and competent authorities, facilitating smoother investigations and more effective risk mitigation measures.

In summary, cooperation with competent authorities is a fundamental aspect of the reporting process outlined in the EU AI Act. It underscores the shared responsibility of providers and authorities in safeguarding AI system safety and reliability, ultimately strengthening AI safety measures and enhancing public confidence in the technology.

Confidentiality and Compliance Guidance

Ensuring confidentiality is paramount in the reporting and investigation of serious incidents under the EU AI Act. Providers must handle sensitive information with utmost care to maintain trust and integrity throughout the process.

The Act imposes strict confidentiality obligations on providers regarding the information obtained during incident reporting and investigation. This includes protecting the privacy of individuals involved, as well as safeguarding proprietary information and trade secrets. Such measures are essential for maintaining confidentiality while upholding regulatory requirements.

To assist providers in fulfilling these obligations, the Commission plays a pivotal role in developing dedicated guidance. This guidance offers practical insights and best practices for compliance, helping providers navigate the complex landscape of confidentiality requirements effectively.

By adhering to these guidelines, providers can ensure the confidentiality of sensitive information while fulfilling their reporting and cooperation obligations under the EU AI Act. This proactive approach not only strengthens AI safety measures but also enhances trust and transparency in the industry, ultimately benefiting both providers and stakeholders alike.

Challenges and Best Practices

Meeting reporting obligations and cooperating with authorities can pose significant challenges for AI system providers. One common challenge is the complexity of incident assessment, especially in determining the severity of incidents and their causal links to AI systems. Additionally, navigating varying reporting timelines and requirements across different jurisdictions can be daunting.

To address these challenges, providers should implement best practices for effective communication and collaboration. Firstly, establishing clear internal protocols for incident identification, assessment, and reporting can streamline the process. Regular training sessions for staff members involved in incident management can enhance their understanding of reporting requirements and ensure compliance.

Moreover, maintaining open lines of communication with competent authorities is essential. Establishing direct points of contact and fostering a cooperative relationship can facilitate timely information exchange and smoother investigations. Providers should also prioritize transparency and honesty in their interactions with authorities, promptly disclosing relevant information and cooperating fully with inquiries.

Adhering to the regulatory framework while embracing proactive measures can mitigate challenges and promote a culture of safety and accountability within the AI industry. By implementing these best practices, providers can navigate reporting obligations and collaboration with authorities effectively, ultimately strengthening AI safety and regulatory compliance.

Conclusion

In conclusion, timely and transparent reporting of serious incidents is paramount in upholding AI safety standards and fostering public trust. Cooperation between AI system providers and competent authorities is essential for swift and effective incident resolution. By adhering to reporting obligations and collaborating closely with authorities, providers can proactively address potential risks and maintain a safe AI environment. These processes not only ensure compliance with the EU AI Act but also demonstrate a commitment to accountability and user safety. Ultimately, by prioritizing timely reporting and cooperation, stakeholders contribute to the establishment of a robust and trustworthy AI ecosystem that benefits society as a whole.

Glossary

Act or EU AI Act: European Union Artificial Intelligence Act

AI: Artificial Intelligence

Board: European Union Artificial Intelligence Board

EU: European Union

SME: Small and Medium-Sized Enterprise

How can we help?

AI & Partners
Amsterdam - London - Singapore

AI & Partners – 'AI That You Can Trust'

At AI & Partners, we're here to help you navigate the complexities of the EU AI Act, so you can focus on what matters—using AI to grow your business. We specialize in guiding companies through compliance with tailored solutions that fit your needs. Why us? Because we combine deep AI expertise with practical, actionable strategies to ensure you stay compliant and responsible, without losing sight of your goals. With our support, you get AI you can trust—safe, accountable, and aligned with the law.

To find out how we can help you, email contact@ai-and-partners.com or visit https://www.ai-and-partners.com.