Openpolicy

January 15, 2025
To: EU AI Office

OpenPolicy comments on
Second Draft General-Purpose AI Code of Practice

Overview

OpenPolicy appreciates the opportunity to provide comments on the General-Purpose AI Code of Practice (GP-AI Code) and commends the drafting committee's efforts to build a framework that prioritizes transparency, systemic risk mitigation, and robust governance for general-purpose AI models. OpenPolicy is a technology company seeking to democratize the ability of innovative companies of all sizes to engage with policymakers and provide feedback on relevant policy deliverables. OpenPolicy is further engaged in emphasizing the need to use AI, innovation, and technology to foster open policymaking and broader use of technology to streamline compliance and governance automation. We anticipate active engagement with the EU AI Office as this code of practice develops and are ready to offer additional input as necessary. We remain eager to support the implementation of the EU AI Act and its objectives. We are committed to supporting the use and implementation of this essential document designed to guide organizations designing, developing, deploying, or using AI systems to help manage AI risks and promote trustworthy and responsible development and use, actively engaging with other EU agencies such as the European Commission, ENISA and other implementing US agencies alongside our participation in the NIST AI Safety Institute Consortium.

Following the EU AI Office's active engagement with proactive Working Groups, government, industry, academia, and the public to understand their needs and develop practical, updated solutions and best practices, we believe that an open and collaborative policymaking dialogue is essential. This approach enhances the implementation of this code of practice best practices and ongoing efforts to ensure a secure AI lifecycle by focusing on the core risks posed by AI, which is critical to fostering an appropriate, updated, and scalable risk management framework. The participation of innovative companies, particularly startups specializing in cutting-edge AI security, cloud, IoT, and safety solutions, significantly contributes to this effort.

Indeed, many technologies used to support the requirements of the EU AI Act referenced in this draft are developed by these innovative companies and startups. These technologies evaluate the measurements, considerations, and development of AI product architecture, deployment strategies, and overarching cybersecurity, data privacy, and secure-by-design perspectives that should guide manufacturers, which OpenPolicy actively collaborates with

these communities.

As advocates for secure, transparent, and trustworthy AI ecosystems, we recognize the significant progress reflected in the current draft. The measures outlined to address systemic risks, promote transparency, and enhance cybersecurity aligns with the foundational principles of responsible AI governance. However, we believe the Code can be further refined to ensure its measures are both comprehensive and actionable. Further, drawing on our expertise and insights from NIST's AI Risk Management Framework and NISA AI Safety Institute Consortium, we believe the GP-AI Code can further enhance its provisions by incorporating dynamic governance tools, lifecycle security measures, and real-time risk management.

This document provides detailed recommendations for specific commitments and measures to enhance the Code's effectiveness while maintaining its focus on fostering innovation and trust.

Governance and Transparency

The GP-AI Code, through Commitment 1, Measure 1.1, appropriately highlights the importance of documenting the AI model's architecture, training processes, and outputs to ensure transparency and accountability. However, this measure does not adequately address the broader environment in which AI models are developed and deployed, where significant risks can originate. To address this gap, we recommend the introduction of an AI Bill of Materials (AI-BOM) as part of the required documentation. An AI-BOM could provide a comprehensive inventory of all components and dependencies associated with the model, including third-party libraries, datasets used during training, and the runtime environments in which the model operates. Specifically, this documentation should log details of the tools and resources integrated during development and deployment, with particular attention to cloud-based and containerized systems. These environments are often subject to unique security challenges, such as tenant isolation failures and supply chain vulnerabilities, that require additional visibility and oversight.

Recent security incidents have underscored the critical need for such transparency. For example, vulnerabilities in open-source libraries and containerized systems have been exploited by attackers to gain unauthorized access, compromise environments, and contaminate artifacts shared across users. An AI-BOM would mitigate these risks by enabling organizations to trace dependencies, identify potential weaknesses, and implement timely remediation strategies. This approach is particularly important for fine-tuned or repurposed models, where modifications may introduce new vulnerabilities. By requiring the creation and maintenance of an AI-BOM, the GP-AI Code would align with broader efforts to address systemic risks and supply chain security concerns. This

enhanced transparency would not only bolster trust in AI systems but also improve their resilience to evolving threats, ensuring they remain secure and reliable throughout their lifecycle.

Lifecycle Risk Management and Continuous Assessment

The GP-AI Code of Practice makes significant strides in addressing risk assessments and systemic risks and establishing verification and governance mechanisms, particularly through Commitments 6, 4, and 15. However, the current framework could be strengthened by expanding its scope to incorporate dynamic risk management and adaptive governance measures that extend beyond pre-deployment phases. While Measure 6.3 focuses on risk assessment before deployment to address the dynamic nature of operational environments, we recommend extending these requirements to explicitly include continuous risk assessments and verification during deployment and post-deployment phases, as outlined in Measure 6.4. The Code would ensure systemic risks are proactively identified and mitigated in real time by mandating ongoing evidence collection and regular evaluations throughout the model lifecycle.

Real-time validation processes should be embedded into Measure 6.4 to complement pre-deployment assessments. Tools that continuously evaluate operational performance and identify systemic vulnerabilities will ensure compliance with evolving risk profiles. This approach is especially critical for addressing threats currently observed, such as adversarial manipulation of model weights and unauthorized exploitation of cloud-based language models. By incorporating these measures into the GP-AI Code, providers can proactively address security concerns while adhering to the principles of continuous compliance and dynamic risk mitigation. These refinements will ensure the Code remains effective in safeguarding general-purpose AI systems against evolving threats.

Operational environments are inherently dynamic, with evolving threat landscapes that present unforeseen risks. Integrating real-time monitoring tools and anomaly detection mechanisms into lifecycle management processes would create a robust safeguard against emerging threats. For example, these systems could detect anomalies indicative of adversarial behavior or unauthorized access, enabling rapid mitigation of vulnerabilities before they are exploited. Such measures align with cybersecurity best practices, which recognize that operational environments are inherently dynamic and frequently subject to unforeseen risks. This approach is critical for maintaining compliance with Article 55 of the AI Act, which emphasizes continuous assessment of systemic risks at the Union level.

In addition, the governance frameworks outlined in Commitments 4 and 15 provide a solid foundation for managing systemic risks. However, relying solely on periodic documentation,

as required under Measures KPI 15.1, and KPI 15.2, may leave gaps in monitoring and compliance, particularly in dynamic environments. We recommend integrating automated tools for real-time monitoring and verification of adherence to safety frameworks, allowing providers to dynamically assess evolving risks, identify discrepancies in documentation, and detect gaps in mitigation strategies. For example, Measure 15.2, which mandates positive adherence to safety frameworks, could be expanded to require periodic automated risk audits throughout the model lifecycle. Such enhancements would provide actionable insights to improve security and governance while enabling the rapid identification of compliance gaps, complementing the Code's objectives by ensuring adaptive governance and reducing response times to emerging threats.

The taxonomy of systemic risks in Commitment 3 plays a vital role in identifying high-risk areas, such as adversarial attacks, large-scale manipulation, and cyber-offensive capabilities. To strengthen this section, the taxonomy could address risks related to AI-driven exploit generation, data poisoning, and the misuse of AI for large-scale disinformation campaigns. Including illustrative examples—such as how automated exploit generation could scale cyberattacks—would enhance understanding and mitigation efforts. Similarly, systemic risks tied to model propensities, such as misalignment or deceptive behavior, should be framed within real-world contexts to underscore their significance and guide providers in implementing effective mitigation strategies.

Further, Measure 3.4 should integrate real-time detection systems to monitor for anomalous behavior, unauthorized disclosures, or adversarial activity to proactively source systemic risks and address vulnerabilities that have already been exploited in the industry, such as the compromise of open-source AI libraries like Ultralytics or adversarial manipulations targeting cloud-hosted AI systems. A comprehensive taxonomy, enriched with specific scenarios and contextual examples, will ensure systemic risks are addressed effectively across diverse deployment contexts.

Strengthening Security Mitigations

Under Commitment 12, the GP-AI Code introduces essential measures to ensure security readiness, including the deployment of Endpoint Detection and Response (EDR) and Intrusion Detection Systems (IDS). While these provisions establish a strong foundation, the Code does not explicitly address the unique challenges posed by cloud-native environments, which are increasingly integral to AI operations. We recommend that the Code expand its focus to explicitly include cloud-based containers and workloads, where vulnerabilities such as tenant isolation failures have been widely observed.

To strengthen these provisions, the Code should mandate the use of detection tools

tailored to containerized environments. These tools must account for the distinct attack vectors that arise in such setups, ensuring effective identification and response to threats. Moreover, the Code could require the correlation of security logs across real-time signals and cloud activities to detect lateral movements—a common tactic used by adversaries to exploit gaps between containerized workloads and underlying infrastructure.

Additionally, logging requirements under Commitment 12 should be expanded to enhance both security and transparency. Specifically, logs should capture inference calls made to models, allowing providers to detect potential misuse or anomalous activity. Similarly, access logs for model weights, as highlighted in Measures 12.3 and 12.4, should track changes to privileged access and any interaction with sensitive components. These logs are critical for forensic investigations and can enable organizations to respond effectively to sophisticated attacks, such as those exploiting vulnerabilities in GPU operators or targeting dependencies in the AI supply chain.

Furthermore, interoperability testing frameworks should be incorporated into Measure 12.4 to ensure that AI models integrate seamlessly with downstream applications while maintaining robust security standards. Commitment 1 could complement this effort by requiring providers to document the lineage of training data, including acquisition methods, preprocessing steps, and provenance verification. This transparency aligns with Article 53(1) of the AI Act and minimizes the risk of incompatibilities and misuse in diverse operational environments.

Enhancing Data Privacy and Secure Practices

Although Commitments 2 (Copyright Policy) and 20 (Documentation) play a crucial role in safeguarding intellectual property and sensitive datasets, they require further specificity to effectively address the growing data security and privacy risks. To strengthen these provisions, we recommend incorporating robust encryption standards, granular access controls, and audit trails to ensure the security of training datasets and model weights. These mechanisms are essential for aligning the Code with global data protection standards and preventing unauthorized access or breaches.

For instance, Measure 20.1 could explicitly mandate the encryption of model weights and related assets, providing an essential safeguard against unauthorized access. Additionally, Measure 2.9, which focuses on mitigating overfitting, should require the logging of training data usage to identify patterns that could inadvertently lead to copyright infringements. This enhancement would align with Article 53(1)(b) of the AI Act, which emphasizes the importance of documentation to enable downstream providers to understand and address associated risks.

Openpolicy

Beyond encryption and access controls, implementing measures to log data lineage and access activities would significantly enhance transparency and accountability. These logs would empower providers to detect unauthorized use of datasets and trace potential vulnerabilities in their AI systems. By providing clear audit trails, organizations can strengthen their ability to identify and remediate risks in a timely and effective manner.

Preparing for Emerging Threats

The GP-AI Code of Practice should consider addressing the growing risks associated with supply chain vulnerabilities and tenant isolation weaknesses, particularly in shared environments. These threats pose significant challenges to the security and integrity of AI systems, especially in AI-as-a-Service contexts where multiple users interact with shared infrastructure. To mitigate these risks, providers should be required to vet all third-party libraries and dependencies used during model development and deployment. This vetting process should include rigorous security audits conducted regularly to identify and address vulnerabilities within supply chain components.

In addition to securing the supply chain, the Code could mandate robust isolation measures to protect against cross-tenant vulnerabilities and data leakage. Tenant isolation techniques, which prevent unauthorized access or data sharing between users in shared environments, are critical for maintaining operational security. These measures must be complemented by periodic testing of sandbox environments, which allow providers to safely test and evaluate new models or updates in a controlled setting. Testing sandbox environments for vulnerabilities ensures that weaknesses are identified and remediated before deployment, reducing the risk of compromise.

These recommendations reflect lessons learned from recent supply chain attacks on widely used AI libraries and software dependencies, where attackers exploited weaknesses in shared resources to gain unauthorized access or disrupt operations. By implementing these measures, the GP-AI Code would proactively address the evolving threat landscape while supporting innovation and scalability. Enhanced supply chain security and robust isolation practices not only protect against emerging threats but also bolster trust and reliability in AI systems across diverse use cases. More broadly additional focus should be given to the risks presented to entire system, environment or architecture from AI, beyond the model risks.

We appreciate your consideration and look forward to our continued collaboration. We remain at your disposal for any further questions,

/s/ Michelle Sahar

Openpolicy

Michelle Sahar
Cybersecurity Policy Director, OpenPolicy

/s/ Dr. Amit Elazari
Dr. Amit Elazari
CEO and Co-Founder of OpenPolicy