

International Guiding Principles and Cybersecurity Guidelines for AI Actors Adopted by G7 and Other Major Economies

CONTRIBUTORS

Kimberly Parry

Karol Piwonski



Julia Anderson

ALERTS

December 11, 2023

During the last weeks of 2023, the international community announced several initiatives designed to establish common standards for AI actors to implement in their development and use of AI. On October 30, 2023, the leaders of the G7 released [International Guiding Principles](#) and a [Code of Conduct](#) for organizations developing and using advanced AI systems. The following month, the cybersecurity authorities of the UK, the U.S., and more than a dozen other major economies published the [Guidelines for Secure AI System Development](#) on November 27, 2023. While not legally binding, these instruments define proposed best practices for developing trustworthy AI, establish an international understanding of the risks and mitigation strategies related to AI systems, and complement national laws regulating AI.

Background Information

The International Guiding Principles and Code of Conduct were developed pursuant to the Hiroshima AI Process. The Hiroshima AI Process was established in May 2023 by the G7 (the U.S., the UK, Canada, France, Germany, Italy, and Japan) and the EU. The purpose of this initiative is to provide guidance for organizations developing AI and to support international cooperation on promoting safe and trustworthy AI. The Hiroshima AI Process complements ongoing discussions within other international forums, such as the Organization for Economic Co-operation and Development (OECD), the Global Partnership on Artificial Intelligence (GPAI), and the UK's AI Safety Summit.

The Guidelines for Secure AI System Development are a follow-up to the UK AI Safety Summit held in London in early November. The Summit brought together international governments, leading AI companies, civil society groups, and experts in research to consider the risks of AI and discuss how those risks can be mitigated through internationally coordinated action.

International Guiding Principles and Code of Conduct for AI

The aim of the International Guiding Principles is to prompt the development of “safe, secure, and trustworthy AI worldwide.” They are directed to AI actors involved in the design, development, deployment, and use of advanced AI systems, which includes foundation models and generative AI systems.

The Guiding Principles can be summarized as follows:

1. Identify, evaluate, and mitigate risks throughout the AI lifecycle (from development to market)
2. Monitor and address vulnerabilities, incidents, risks, and misuse after deployment
3. Increase accountability via transparency about advanced AI systems' capabilities, limitations, and domains of use

4. Collaboratively share information and report incidents with stakeholders across the AI lifecycle (e.g., industry, governments, civil society, and academia)
5. Enact and disclose AI governance and risk management policies, grounded in a risk-based approach
6. Invest in and implement robust security controls
7. Implement content authentication and provenance mechanisms to allow users to identify AI-generated content
8. Prioritize research on risk mitigation and prioritize investment in mitigation measures
9. Prioritize the development of advanced AI systems to address the greatest societal challenges (such as climate change, global health, and education)
10. Advance the development and use of international technical standards
11. Take appropriate measures to manage data quality and to protect personal data and intellectual property

The Code of Conduct builds upon the Guiding Principles and provides recommendations for specific actions organizations can take to implement them when designing, developing, and using advanced AI systems to minimize the risks they pose. It encourages organizations to take a risk-based approach while governments develop more detailed governance and regulatory approaches.

The G7 leaders have called on organizations developing advanced AI systems to commit to the application of the Guiding Principles and the Code of Conduct. To ensure that they remain fit for purpose, both documents will continue to be reviewed and updated as necessary, taking into account ongoing consultations and the latest developments in advanced AI systems.

Guidelines for Secure AI System Development

The Guidelines for Secure AI System Development aim to ensure that AI systems are designed, developed, and deployed securely. The Guidelines set forth the first global, common understanding of cyber risks and mitigation strategies specific to AI systems. They are targeted towards providers of AI systems, though all stakeholders are urged to consider them to make informed decisions about the design, deployment, and operation of their AI systems.

The guidelines follow a “secure by default” approach, which prioritizes taking ownership of security outcomes for customers, embracing radical transparency and accountability, and building organizational structure and leadership so secure design is a top business priority.

The guidelines are grouped into four key areas: secure design, secure development, secure deployment, and secure operation and maintenance. The guidelines pertaining to each area are summarized below.

Secure Design

At the design stage of developing an AI system, developers should understand and consider risks to an AI system alongside other design choices. Specifically, the guidelines recommend that i) system owners and senior leaders understand threats to AI systems and how to mitigate them, ii) there is a holistic system in place to assess threats to an AI system as part of the risk management process, iii) the appropriateness of any AI-specific design choices are analyzed with security in mind in addition to functional and other considerations, and iv) the choice of AI model is informed by a consideration of the threats to that model, which may require reassessment as AI security research advances.

Secure Development

At the development stage, developers should track security-relevant data and set up security procedures to manage risks created as an AI system is being developed. Specifically, the guidelines recommend i) setting security standards for sourcing software and hardware products from outside suppliers for use in AI systems, ii) securely managing AI-related assets created in the process of developing a system such that the assets can be secured and restored in the event of compromise by attackers, iii) capturing security-relevant data in the development of an AI system like the sources of training data and retention time, and iv) managing risks created by engineering decisions that fall short of best practices to achieve short-term results.

Secure Deployment

At the deployment stage, the guidelines emphasize proactive threat detection and security measures at every stage where an AI system connects with end users. They recommend i) the application of

good infrastructure security principles (e.g., access controls) to the infrastructure used in every part of a system's lifecycle, ii) protection of an AI model and data through the implementation of cybersecurity best practices and controls on the query interface that detect and prevent attempts to access confidential information, iii) preparation of incident management procedures addressing a wide range of scenarios and regular reassessment of such procedures, iv) releasing new AI systems only after subjecting them to effective security evaluations, and v) providing guidance to users on the most secure way to use an AI system.

Secure Operation and Maintenance

Finally, at the secure operation and maintenance stage, the guidelines emphasize comprehensive oversight of an AI system and transparency regarding updates and vulnerabilities. Specifically, they recommend i) comprehensive oversight of the outputs of an AI system to track changes that may affect the security of the system, ii) oversight of the inputs to a system to enable mitigation measures in the event of compromise, iii) updated procedures that clearly showcase when changes to a model or its data inputs have been made, and iv) participation in information-sharing communities to identify and address system vulnerabilities and share best practices.

Next Steps

The Hiroshima AI Process and the AI Safety Summit are part of a global attempt to establish mechanisms of international cooperation regarding the development and use of AI systems and to ensure the safety and security of those systems. They complement other mechanisms, such as the U.S. Executive Order on the Safe, Secure, and Trustworthy Development and Use of AI, which we cover [here](#), and the upcoming EU AI Act, which we cover [here](#) and [here](#). G7 leaders have also instructed relevant ministers to develop a "Hiroshima AI Process Comprehensive Policy Framework" by the end of 2023 in cooperation with GPAI and OECD.

The guiding documents are not legally binding and therefore do not mandate companies to make any changes in the way AI systems are designed, developed, deployed, or used. However, companies may want to use the G7 Guiding Principles, G7 Code of Conduct, and Guidelines for Secure AI System Development as a checklist of best practices to consider in benchmarking their AI programs or to implement in their development of new AI systems.

Wilson Sonsini Goodrich & Rosati routinely helps companies navigate complex AI, privacy, and data security issues. For more information or advice concerning your development, compliance, and commercialization efforts related to AI and machine learning, please contact [Maneesha Mithal](#), [Cédric Burton](#), [Laura De Boel](#), [Scott McKinney](#), [Barath Chari](#), or any member of the firm's [privacy and cybersecurity practice](#) or [artificial intelligence and machine learning working group](#).

Kimberly Parry, Karol Piwonski, and Julia Anderson contributed to the preparation of this Wilson Sonsini Alert.