MARCH 2024

Advancing the Hiroshima AI Process Code of Conduct under the 2024 Italian G7 Presidency

Timeline and Recommendations

By Gregory C. Allen and Georgia Adamson

Introduction

Over the past five years, the G7 and its member states have made significant progress in specifying ethical principles for AI. Advancing global AI governance in 2024 demands translating these high-level principles into concrete practices for AI developers and government organizations. Rather than address this challenge in isolation, G7 leaders stated in their **2023 Leaders' Communiqué** they would work together and with others to "advance international discussions on inclusive artificial intelligence (AI) governance and interoperability to achieve our common vision and goal of trustworthy AI, in line with our shared democratic values." The March **2024 Industry, Technology and Digital Ministerial declaration** recently reaffirmed this commitment, stressing "the importance of international discussions on AI governance and interoperability" with like-minded partners and developing countries.

The work of the G7 has already had a tangible impact in productively shaping the domestic AI regulatory approach of multiple G7 members, including the United States (through the 2023 AI executive order), Japan (the updated AI regulatory guidance), and the European Union (the AI Act). In conversations with CSIS, government officials from each of these members stated that the G7 was substantive and helpful to their regulatory efforts. Notably, members of the European Parliament involved in drafting the EU AI Act told CSIS that some sections of the act were directly inspired by the **Hiroshima AI Process (HAIP) Code of Conduct** the G7 published in October 2023.

CSIS | CENTER FOR STRATEGIC & INTERNATIONAL STUDIES

One such section of the EU AI Act is **Article 52**, which outlines transparency obligations for Global Partnership on Artificial Intelligence (GPAI) models above a computational threshold of 10^{25} floating point operations per second (FLOPS). According to Article 52, the European Union's AI Office shall "encourage and facilitate the drawing up of codes of practice at Union level as an element to contribute to the proper application of this Regulation, taking into account international approaches." Industry developers are directed to coregulate these codes of practice by collaborating directly with the European Commission on this issue.

Members of the European Parliament told CSIS they believe the HAIP Code of Conduct is the best available starting point for what will ultimately become the EU AI Office's officially recognized codes of practice for certain GPAI models. Parliament members stated their support for maturing the G7's code of conduct to a degree that meets EU requirements, such that providers of GPAI models who demonstrate compliance with the G7's HAIP Code of Conduct would enjoy a "presumption of conformity" to the EU AI Act's GPAI codes of practice.

The European Union, while it has the sole final decision over the AI Act's implementation, should be commended for its demonstrated commitment to interoperability, openness to international input during the codes of practice design process, and willingness to take advantage of the work already done under the Japanese 2023 G7 presidency. Maturing the code of conduct to confer a presumption of conformity to the EU AI Act would not only help the European Commission develop a key piece of regulation for GPAI models but also offer an opportunity for other G7 members to help shape language that will set an informative precedent for how global AI regulations are designed and implemented. Moreover, aligning the code of conduct to the European Union's codes of practice would be hugely impactful for the G7's **stated** goals of establishing interoperable AI regulations and broadening the dissemination of the code.

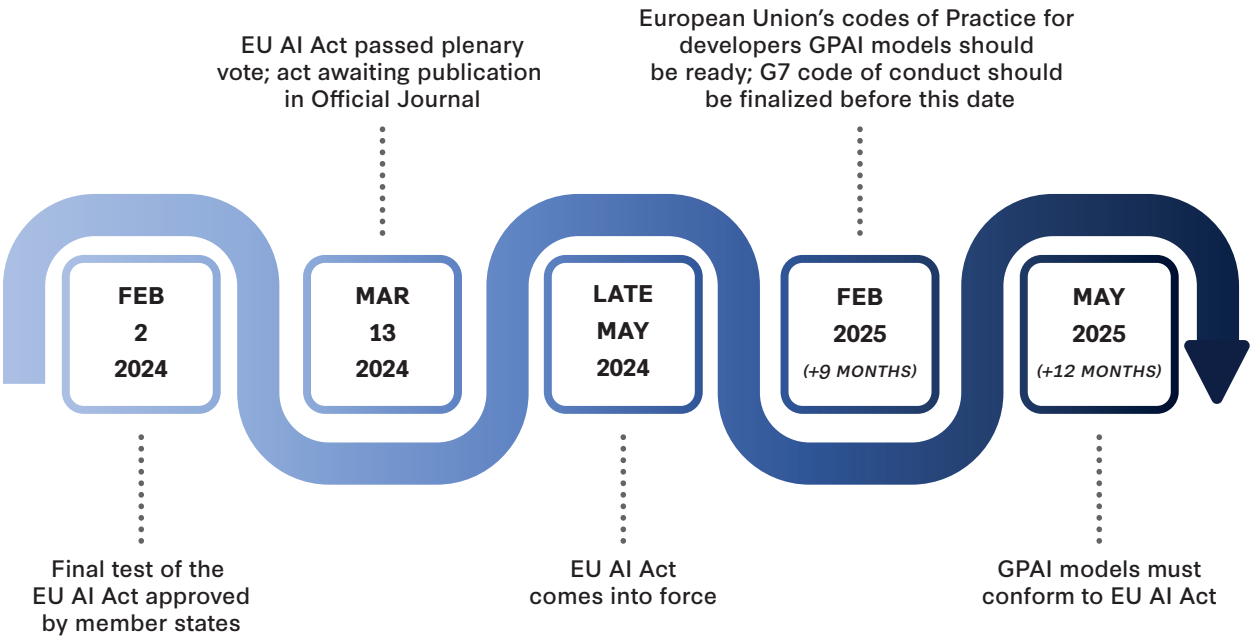
Even if the European Commission does not ultimately use the code of conduct to inform or stand in for the European Union's official codes of practice, the G7 would still achieve a substantially more mature code by aiming to meet this ambitious goal. In its current state, the code of conduct's 11 high-level principles are too vague for AI developers to subscribe to or implement without further specification. Translating these principles into monitorable and enforceable tasks for AI developers would therefore be a meaningful achievement alone. This process will require extensive input from the private sector to determine what is relevant, technically feasible, and reflective of existing industry best practices in AI governance. Cooperation with and public endorsement from a diverse group of leading international AI firms will also be needed to help to demonstrate the code's building momentum within the AI industry. While maturing the code of conduct through these avenues and more would already fulfil one of the G7's stated priorities for 2024, the G7 could go further by aligning the code of conduct with the European Union's codes of practice to achieve its broader goals of enhancing regulatory interoperability and widening the adoption of the code of conduct.

Timeline

The expected **timeline** for EU AI Act implementation suggests that a draft version of the GPAI codes of practice must be ready by February 2025 (see Figure 1). Comparing the EU and G7 timelines highlights that the G7 code of conduct should be substantially elaborated by February 2025 if the European Union is to consider the code of conduct as a presumption of conformity to the European Union's code of

practice on GPAI models when the regulations take effect. This gives the G7 approximately 11 months from the writing of this white paper to mature the code of conduct ahead of the European Union's February 2025 deadline. While the HAIP Code of Conduct could inform a later, updated EU AI codes of practice, meeting the February 2025 deadline would certainly have greater impact.

Figure 1: Expected EU AI Act Implementation Time Frame



Source: Authors' analysis based on multiple sources.

The Italian G7 presidency therefore comes at a critical window of opportunity to substantially advance the HAIP and the interoperability of allied AI regulatory frameworks. The Italian presidency recently **restated** the G7's commitment to maturing the HAIP and the code of conduct at the Industry, Technology, and Digital Ministerial Meeting in Verona and Trento on March 14 and 15, 2024. The ministerial declaration from this meeting **promises** to advance the HAIP outcomes on page 10, "including through expanding support and awareness among key partners and organisations, as well as increasing their involvement, as appropriate." Specifically, per Annex 3, the G7 **will** develop monitoring mechanisms for assessing organizations' voluntary implementation of the code of conduct, increase stakeholder engagement in maturing the code of conduct and its adoption, and collaborate with the Organization for Economic Cooperation and Development (OECD), UNESCO, and GPAI on challenges posed by AI systems. Progress on these issues will be **presented** at the leaders' summit in June.

Recommendations

Given the G7's 2024 agenda for the HAIP and the European Commission's deadline of February 2025 for developing the GPAI codes of practice, the Italian G7 presidency should strive to accomplish the following this year:

1. **Continue working on the HAIP Code of Conduct after the March 15 ministerial meeting, including a second digital and industry ministerial meeting toward the end of 2024.**

Developing the code of conduct will take substantive effort from both G7 members and AI developers in the private sector and academia. CSIS supports the Italian presidency's year-round approach to maturing the HAIP ending, with a second ministerial meeting showcasing the G7's work toward the end of 2024.

2. **Develop the HAIP Code of Conduct's monitoring and evaluation mechanisms through the OECD**. While the G7 has done excellent work on advancing AI governance principles thus far, its lack of permanent staff makes it challenging to continue this work alone in the long-term. The March Industry, Technology, and Digital Ministerial declaration indicates that the OECD will provide this missing institutional support to the G7 on the Hiroshima AI Process, continuing a long-standing **partnership** between the two fora.

   The OECD has a strong record of **developing** monitoring and evaluation mechanisms, which the G7 should draw on to meaningfully mature the code of conduct in 2024. Japan chairs this year's OECD's Ministerial Council, offering continuity between the G7 Hiroshima Process, which Japanese government officials continue to staff and support, and the OECD. Developing the code of conduct through the Ministerial Council and other OECD channels, including the Global Partnership on AI, will offer essential capacity and expertise to the 2024 G7.

   Close collaboration with the OECD would also offer a longer-term path to expand the HAIP beyond the G7, first to like-minded democratic market-driven economies among the OECD countries. The OECD could also serve as a first step in further expansion to a wider set of non-OECD countries through the OECD-hosted Global Partnership on AI, which includes countries such as Brazil and India.

3. **Gather input from private sector AI organizations on the code of conduct through requests for comment and formalized convenings of relevant stakeholders**. In a **speech** to the Italian government's Digital Transition Committee in November 2023, Prime Minister Giorgia Meloni stated that establishing AI governance guidelines "does not mean working against companies" but instead "engaging in dialogue" between the public and private sector. Dialogue between relevant stakeholders is indeed essential for ensuring the code of conduct is translated into specific actions that are technically feasible and reflect the rapidly changing capabilities of advanced AI systems. To ensure that this dialogue remains consistent and constructive, input from private sector AI developers should be organized through specific initiatives such as requests for comment, working groups, and feasibility studies.

With these recommendation in place, CSIS hopes the following occur by the end of the Italian G7 presidency:

- G7 members and EU AI Act officials expand, finalize, and accept the HAIP Code of Conduct language.
- A substantial number of companies support the drafting of and publicly subscribe to the code of conduct.
- The OECD serves as the principal forum in which to develop monitoring and transparency functions of the code of conduct as necessary.
- By late 2024 or early 2025, there is a clear path to expanding the code of conduct beyond the G7 and OECD forums, including non-G7 and OECD countries potentially through the OECD's Global Partnership on AI forum. ■

Gregory C. Allen *is director of the Wadhwani Center for AI and Advanced Technologies and senior fellow with the Strategic Technologies Program at the Center for Strategic and International Studies (CSIS) in Washington, D.C.* ***Georgia Adamson*** *is a research assistant with the Wadhwani Center for AI and Advanced Technologies at CSIS.*

This report is made possible through the generous support of the Japanese government.

Appendix A: CSIS's Work on the HAIP Code of Conduct

Recognizing the importance of developing the HAIP in 2024, CSIS is hosting events and publishing papers to help support the G7 workstream on this issue. Here is a list of the Center's past and upcoming work on the HAIP:

MARCH 2024

- **Workshop 1: HAIP Code of Conduct: Introduction and Fulfilment Structures**
  Date: March 5
  Description: This first of two private workshops gathered representatives of major tech companies, academic institutions, and G7 countries' governments to review the HAIP, identify next steps for the code of conduct, and consider technical feasibility for AI developers.

- **Workshop 2: The HAIP: Next Steps and Future Work Streams**
  Date: March 21
  Description: This second of two workshops reconvened participants to consider proposed workstreams for advancing the HAIP Code of Conduct under the Italian G7 Presidency, and debated directions for broader international cooperation on the HAIP in 2024.

APRIL 2024

- **Public event**
  Date: TBD
  Description: Yoichi Ida, assistant vice minister for international affairs at the Japanese Ministry of Internal Affairs and Communication, will join CSIS Wadhwani Center director Gregory Allen to discuss global AI governance efforts and the G7 HAIP.

EARLY SUMMER 2024

- **White paper**
  Date: TBD
  Description: This will include further recommendations for maturing the Hiroshima Process Code of Conduct in 2021.

Appendix B: CSIS Summary of the HAIP Code of Conduct

On October 30, 2023, G7 leaders announced the Hiroshima Process Code of Conduct, a list of high-level voluntary guidelines for organizations to develop safer and more transparent advanced AI. The code defines advanced AI systems as frontier foundation models and generative AI systems. Organizations that may voluntarily commit to the code of conduct include entities from industry, academia, civil society, or the public sector.

The code of conduct requires developers who have subscribed to the code to do the following:

1. IDENTIFY, EVALUATE, AND MITIGATE RISKS OF AI ACROSS THE DEVELOPMENT LIFE CYCLE.

- Develop and employ internal and independent external testing measures, such as red teaming, to address systems' risks and vulnerabilities and to promote safety, security, and trust across the AI life cycle.
- Where appropriate, closely monitor and respond to risks relating to human and environmental health and safety, human rights, bias and discrimination, and democracy.

2. IDENTIFY AND MITIGATE VULNERABILITIES, AS WELL AS INCIDENTS OR PATTERNS OF MISUSE, AFTER DEPLOYMENT.

- Monitor and address cases of misuse and emerging risks in deployed AI systems.
- Adopt mechanisms to incentivize the reporting of vulnerabilities or misuse cases by users and other third parties, including through bounty systems, contests, or prizes.
- Maintain up-to-date documentation and reports of all recorded vulnerabilities and incidents of misuse.

3. PUBLICLY REPORT AI SYSTEMS' CAPABILITIES, LIMITATIONS, AND DOMAINS OF USE.

- Publish up-to-date transparency reports regarding AI systems' domains of appropriate or inappropriate use, capacities and limitations, and risks to safety and society, as well as the results of safety evaluations conducted.

4. WORK TOWARD RESPONSIBLE INFORMATION SHARING AMONG AI DEVELOPMENT STAKEHOLDERS.

- Where appropriate, collaborate with industry, government, civil society, and academia to share information for the purpose of establishing common standards and practices on AI safety, security, and trustworthiness. Examples include evaluation reports, security and safety risks, dangerous capabilities, and actions by AI actors to evade safeguards.

5. DEVELOP, IMPLEMENT, AND DISCLOSE AI RISK MITIGATION POLICIES.

- Following a risk-based approach, establish, update, and disclose AI risk mitigation policies and, where appropriate, privacy policies, including for personal data, user prompts, and advanced AI outputs.

6. INVEST IN AND IMPLEMENT STRONG SECURITY MEASURES.

- Invest in physical security, cybersecurity, and insider threat safeguards, including securing model weights, algorithms, servers, and data sets for information security and cyber/physical access controls.
- Establish a robust insider threat detection mechanism to protect high-value intellectual property and trade secrets.

7. DEVELOP AND DEPLOY PROVENANCE AND AUTHENTICATION MECHANISMS FOR USERS TO IDENTIFY AI.

- Implement provenance mechanisms for AI-created content, including the identity of the service or model that created the content.
- Develop content authentication tools such as watermarks that notify users of AI-generated content and implement disclaimers to alert users when interacting with AI systems.

8. PRIORITIZE RESEARCH FOR AI RISK MITIGATION.

- Invest in research that promotes democratic values; human rights; protection of children and vulnerable groups; intellectual property rights; and privacy, bias, misinformation, and disinformation.
- Share risk mitigation research between AI development organizations.

9. PRIORITIZE AI DEVELOPMENT TO ADDRESS PRESSING GLOBAL CHALLENGES.

- Support progress of UN Sustainable Development Goals and encourage AI development for pressing global challenges.
- Support digital literacy initiatives to promote education and public training on AI systems.
- Collaborate with civil society and community groups to achieve these goals.

10. ADVANCE AND ADOPT INTERNATIONAL TECHNICAL STANDARDS.

- Work with standards development organizations to develop and implement interoperable international technical standards, including for watermarking, public reporting, and cybersecurity measures.

11. IMPLEMENT APPROPRIATE DATA INPUT MEASURES AND PROTECTIONS FOR PERSONAL DATA AND INTELLECTUAL PROPERTY.

- Manage data inputs and implement safeguards to ensure data quality, privacy preservation, and intellectual property rights.