参考資料６－２

（仮訳）

高度な AI システムを開発する組織向けの広島プロセス国際指針

高度な AI システムを開発する組織向けの広島プロセス国際指針は、安全、安心、信頼できる AI を世界に普及させることを目的とし、最先端の基盤モデル及び生成 AI システムを含む、最も高度な AI システム（以下、「高度な AI システム」）を開発・利用する組織のための指針を提供するものである。組織には、学術界、市民社会、民間部門、公共部門の主体等が含まれる。

この非網羅的な指針のリストは、高度な AI システムの近年の発展に応じて、既存の OECD AI 原則を基礎とする生きた文書として議論され、精緻化されるものであり、このような AI 技術によってもたらされる利益を捉え、リスクと課題に対処することを補助することを意図している。これらの原則は、高度な AI システムの設計、開発、導入、利用をカバーするために、必要に応じて、すべての AI 関係者に適用されるべきである。

我々は、他の国々や、学術界、ビジネス界、市民社会等のより広範なステークホルダーの意見を取り入れながら、包括的な政策枠組みの一部としてこれらの原則をさらに発展させていくことを期待している。

また、我々は、以下の指針に基づき、高度な AI システムを開発する組織向けの国際的な行動規範を精緻化することへのコミットメントを改めて表明する。

国や地域によって、異なる方法でこれらの指針を実践するために、独自のアプローチをとることができる。

我々は、各国政府がより永続的かつ／又は詳細なガバナンスと規制のアプローチを策定する一方で、リスクベースのアプローチに沿って、組織が、他の関係するステークホルダーと協議の上、これらの行動に従うことを求める。また、我々は、OECD や GPAI、その他ステークホルダーと協議の上、組織がこれらの行動の実施について説明責任を維持するためのモニタリングツール及びメカニズムを導入するための提案を作成することにコミットする。我々は、ベストプラクティスを提供することによって、我々が開発しようとする効果的なモニタリングメカニズムの開発を支援することを組織に奨励する。

イノベーションの機会を活用する一方で、組織は、高度な AI システムの設計、開発、導入において、法の支配、人権、適正手続き、多様性、公平性、無差別、民主主義、人間中心主義を尊重すべきである。

組織は、民主主義の価値観を損ない、個人や地域社会に特に有害であり、テロリズムを助長し、犯罪的な悪用を可能にし、安全、セキュリティ、人権に重大なリスクをもたらすような方法で、高度な AI システムを開発・導入すべきではなく、そのようなやり方は容認できない。

国家は、人権が十分に尊重され保護されるよう、国際人権法上の義務を遵守しなければならない。一方、民間部門の活動は、国連「ビジネスと人権に関する指導原則」や OECD「多国籍企業行動指針」等の国際的枠組みに沿ったものでなければならない。

具体的には、各組織に対し、リスクに見合った形で、以下の原則を遵守するよう求める。

1. ***AI ライフサイクル全体にわたるリスクを特定、評価、軽減するために、高度な AI システムの開発全体を通じて、その導入前及び市場投入前も含め、適切な措置を講じる***

これには、レッドチーム等の様々な手法を組み合わせて、多様な内部テスト手段や独立した外部テスト手段を採用することや、特定されたリスクや脆弱性に対処するための適切な緩和策を実施することが含まれる。テストと緩和策は、例えば、システムが不合理なリスクをもたらさないように、ライフサイクル全体を通じてシステムの信頼性、安全性、セキュリティの確保を目指すべきである。このようなテストを支援するために、開発者は、データセット、プロセス、システム開発中に行われた意思決定に関連して、トレーサビリティを可能にするよう努めるべきである。

2. ***市場投入を含む導入後、脆弱性、及び必要に応じて悪用されたインシデントやパターンを特定し、緩和する***

組織は、リスクレベルに見合った適切なタイミングで、AI システムを意図したとおりに使用し、導入後の脆弱性、インシデント、新たなリスク、悪用を監視し、それらに対処するための適切な措置を講じるべきである。組織は、例えば、導入後に第三者やユーザーが問題や脆弱性を発見し報告することを促進することの検討が奨励される。組織はさらに、他の利害関係者と協力して、報告されたインシデントの適切な文書化を維持し、特定されたリスクと脆弱性を軽減することが奨励される。適切な場合には、脆弱性を報告する仕組みは、多様な利害関係者が利用できるものでなければならない。

3. ***高度な AI システムの能力、限界、適切・不適切な使用領域を公表し、十分な透明性の確保を支援することで、アカウンタビリティの向上に貢献する***

これには、高度な AI システムの重要な新規公表全てについて、有意義な情報を含む透明性報告書を公表することが含まれるべきである。

組織は、適切かつ関連性のある導入者及び利用者がモデル／システムのアウトプットを解釈し、利用者がそれを適切に利用できるようにするために、透明性報告書内の情報を十分に明確で理解可能なものにすべきである。また、透明性報告書は、強固な文書化プロセスによってサポートされ、提供されるべきである。

4. ***産業界、政府、市民社会、学界を含む、高度な AI システムを開発する組織間での責任ある情報共有とインシデントの報告に向けて取り組む***

これには、評価報告書、セキュリティや安全性のリスク、危険な意図的又は意図しない能力、AI のライフサイクル全体にわたるセーフガードを回避しようとする AI 関係者の試みに関する情報等を含むが、これらに限定されない、適切な情報の責任ある共有が含まれる。

5. ***特に高度な AI システム開発者に向けた、個人情報保護方針及び緩和策を含む、リスクベースのアプローチに基づく AI ガバナンス及びリスク管理方針を策定し、実施し、開示する***

これには、個人データ、ユーザープロンプト、高度な AI システムのアウトプットを含め、適切な場合にはプライバシーポリシーを開示することが含まれる。組織は、リスクベースのアプローチに従って、AI ガバナンス方針とこれらの方針を実践するための組織的メカニズムを確立し、開示することが期待される。これには、AI のライフサイクルを通じて実行可能な場合には、リスクを評価し、軽減するための説明責任とガバナンス・プロセスが含まれるべきである。

6. ***AI のライフサイクル全体にわたり、物理的セキュリティ、サイバーセキュリティ、内部脅威に対する安全対策を含む、強固なセキュリティ管理に投資し、実施する***

これには、情報セキュリティのための運用上のセキュリティ対策や、適切なサイバー／物理的アクセス制御等を通じて、モデルの重み、アルゴリズム、サーバー、データセットを保護することが含まれる。

7. ***技術的に可能な場合は、電子透かしやその他の技術等、ユーザーが AI が生成したコンテンツを識別できるようにするための、信頼できるコンテンツ認証及び来歴のメカニズムを開発し、導入する***

これには、適切かつ技術的に実現可能な場合、組織の高度な AI システムで作成されたコンテンツのコンテンツ認証及び来歴メカニズムが含まれる。来歴データには、コンテンツを作成したサービス又はモデルの識別子を含めるべきであるが、ユーザー情報を含める必要はない。組織はまた、透かし等を通じて、特定のコンテンツが高度な AI システムで作成されたかどうかをユーザーが判断できるツールや API の開発に努めるべきである。

組織はさらに、可能かつ適切な場合には、利用者が AI システムと相互作用していることを知ることができるよう、ラベリングや免責事項の表示等、その他の仕組みを導入することが奨励される。

8. 社会的、安全、セキュリティ上のリスクを軽減するための研究を優先し、効果的な軽減策への投資を優先する

これには、AI の安全性、セキュリティ、信頼性の向上を支援し、主要なリスクに対処する研究の実施、協力、投資及び適切な緩和ツールの開発への投資が含まれる。

9. 世界の最大の課題、特に気候危機、世界保健、教育等（ただしこれらに限定されない）に対処するため、高度な AI システムの開発を優先する

これらの取り組みは、国連の持続可能な開発目標の進捗を支援し、グローバルな利益のため AI の開発を奨励するために行われる。

組織は、信頼できる人間中心の AI の責任あるスチュワードシップを優先し、また、デジタルリテラシーのイニシアティブを支援すべきである。

10. 国際的な技術規格の開発を推進し、適切な場合にはその採用を推進する

これには、電子透かしを含む国際的な技術標準とベストプラクティスの開発に貢献し、適切な場合にはそれを利用し、標準開発組織（SDO）と協力することが含まれる。

11. 適切なデータインプット対策を実施し、個人データ及び知的財産を保護する

組織は、有害な偏見バイアスを軽減するために、訓練データやデータ収集など、データの質を管理するための適切な措置を講じることが奨励される。

訓練用データセットの適切な透明性も支援されるべきであり、組織は適用される法的枠組みを遵守すべきである。

Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI System

The Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems aims to promote safe, secure, and trustworthy AI worldwide and will provide guidance for organizations developing and using the most advanced AI systems, including the most advanced foundation models and generative AI systems (henceforth "advanced AI systems"). Organizations may include, among others, entities from academia, civil society, the private sector, and the public sector.

This non-exhaustive list of guiding principles is discussed and elaborated as a living document to build on the existing OECD AI Principles in response to recent developments in advanced AI systems and are meant to help seize the benefits and address the risks and challenges brought by these technologies. These principles should apply to all AI actors, when and as applicable to cover the design, development, deployment and use of advanced AI systems.

We look forward to developing these principles further as part of the comprehensive policy framework, with input from other nations and wider stakeholders in academia, business and civil society.

We also reiterate our commitment to elaborate an international code of conduct for organizations developing advanced AI systems based on the guiding principles below.

Different jurisdictions may take their own unique approaches to implementing these guiding principles in different ways.

We call on organizations in consultation with other relevant stakeholders to follow these actions, in line with a risk-based approach, while governments develop more enduring and/or detailed governance and regulatory approaches. We also commit to develop proposals, in consultation with the OECD, GPAI and other stakeholders, to introduce monitoring tools and mechanisms to help organizations stay accountable for the implementation of these actions. We encourage organizations to support the development of effective monitoring mechanisms, which we may explore to develop, by contributing best practices.

While harnessing the opportunities of innovation, organizations should respect the rule of law, human rights, due process, diversity, fairness and non-discrimination, democracy, and human-centricity, in the design, development and deployment of advanced AI systems.

Organizations should not develop or deploy advanced AI systems in a way that undermines democratic values, are particularly harmful to individuals or communities, facilitate terrorism, enable criminal misuse, or pose substantial risks to safety, security, and human rights, and are thus not acceptable.

States must abide by their obligations under international human rights law to promote that human rights are fully respected and protected, while private sector activities should be in line with international frameworks such as the United Nations Guiding Principles on Business and Human Rights and the OECD Guidelines for Multinational Enterprises.

Specifically, we call on organizations to abide by the following principles, commensurate to the risks:

1. ***Take appropriate measures throughout the development of advanced AI systems, including prior to and throughout their deployment and placement on the market, to identify, evaluate, and mitigate risks across the AI lifecycle.***

This includes employing diverse internal and independent external testing measures, through a combination of methods such as red-teaming, and implementing appropriate mitigation to address identified risks and vulnerabilities. Testing and mitigation measures should for example, seek to ensure the trustworthiness, safety and security of systems throughout their entire lifecycle so that they do not pose unreasonable risks. In support of such testing, developers should seek to enable traceability, in relation to datasets, processes, and decisions made during system development.

2. ***Identify and mitigate vulnerabilities, and, where appropriate, incidents and patterns of misuse, after deployment including placement on the market.***

Organizations should use, as and when appropriate commensurate to the level of risk, AI systems as intended and monitor for vulnerabilities, incidents, emerging risks and misuse after deployment, and take appropriate action to address these. Organizations are encouraged to

consider, for example, facilitating third-party and user discovery and reporting of issues and vulnerabilities after deployment. Organizations are further encouraged to maintain appropriate documentation of reported incidents and to mitigate the identified risks and vulnerabilities, in collaboration with other stakeholders. Mechanisms to report vulnerabilities, where appropriate, should be accessible to a diverse set of stakeholders.

3. ***Publicly report advanced AI systems' capabilities, limitations and domains of appropriate and inappropriate use, to support ensuring sufficient transparency, thereby contributing to increase accountability.***

This should include publishing transparency reports containing meaningful information for all new significant releases of advanced AI systems.

Organizations should make the information in the transparency reports sufficiently clear and understandable to enable deployers and users as appropriate and relevant to interpret the model/system's output and to enable users to use it appropriately, and that transparency reporting should be supported and informed by robust documentation processes.

4. ***Work towards responsible information sharing and reporting of incidents among organizations developing advanced AI systems including with industry, governments, civil society, and academia.***

This includes responsibly sharing information, as appropriate, including, but not limited to evaluation reports, information on security and safety risks, dangerous intended or unintended capabilities, and attempts by AI actors to circumvent safeguards across the AI lifecycle.

5. ***Develop, implement and disclose AI governance and risk management policies, grounded in a risk-based approach – including privacy policies, and mitigation measures, in particular for organizations developing advanced AI systems.***

This includes disclosing where appropriate privacy policies, including for personal data, user prompts and advanced AI system outputs. Organizations are expected to establish and disclose their AI governance policies and organizational mechanisms to implement these policies in accordance with a risk-based approach. This should include accountability and governance processes to evaluate and mitigate risks, where feasible throughout the AI lifecycle.

6. ***Invest in and implement robust security controls, including physical security, cybersecurity and insider threat safeguards across the AI lifecycle.***

These may include securing model weights and algorithms, servers, and datasets, such as through operational security measures for information security and appropriate cyber/physical access controls.

7. ***Develop and deploy reliable content authentication and provenance mechanisms, where technically feasible, such as watermarking or other techniques to enable users to identify AI-generated content***

This includes, where appropriate and technically feasible, content authentication such provenance mechanisms for content created with an organization's advanced AI system. The provenance data should include an identifier of the service or model that created the content, but need not include user information. Organizations should also endeavor to develop tools or APIs to allow users to determine if particular content was created with their advanced AI system such as via watermarks.

Organizations are further encouraged to implement other mechanisms such as labeling or disclaimers to enable users, where possible and appropriate, to know when they are interacting with an AI system.

8. ***Prioritize research to mitigate societal, safety and security risks and prioritize investment in effective mitigation measures.***

This includes conducting, collaborating on and investing in research that supports the advancement of AI safety, security and trust, and addressing key risks, as well as investing in developing appropriate mitigation tools.

9. ***Prioritize the development of advanced AI systems to address the world's greatest challenges, notably but not limited to the climate crisis, global health and education***

These efforts are undertaken in support of progress on the United Nations Sustainable

Development Goals, and to encourage AI development for global benefit.

Organizations should prioritize responsible stewardship of trustworthy and human-centric AI and also support digital literacy initiatives.

10. Advance the development of and, where appropriate, adoption of international technical standards

This includes contributing to the development and, where appropriate, use of international technical standards and best practices, including for watermarking, and working with Standards Development Organizations (SDOs).

11. Implement appropriate data input measures and protections for personal data and intellectual property

Organizations are encouraged to take appropriate measures to manage data quality, including training data and data collection, to mitigate against harmful biases.

Appropriate transparency of training datasets should also be supported and organizations should comply with applicable legal frameworks.