



高度な AI システムを開発する組織向けの 広島プロセス国際行動規範

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行動規範は、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に基づき、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」）を開発する組織による行動のための自主的な手引きを提供するものである。

組織は、リスクベースのアプローチに沿って、これらの措置に従うべきである。

本行動規範を支持する組織には、学术界、市民社会、民間部門、公共部門の主体等が含まれる。

この非網羅的な行動リストは、高度な AI システムの近年の発展に対応して、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。組織は、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、ライフサイクルのすべての段階にこれらの行動を適用すべきである。

この文書は、この急速に発展する技術に対して、目的に適合し、かつ対応可能であることを確保するため、必要に応じて、継続中の包摂的なマルチステークホルダー協議を含めて、レビューされ、更新される。

国や地域によって、異なる方法でこれらの行動を実践するために、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。

さらに、これらの行動の実践とAI開発における責任ある、説明可能な手法を促進するため、自己評価メカニズムを含む内部的なAIガバナンス構造と方針を設定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、配備において、法の支配、人権、デュー・プロセス、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、犯罪的な悪用を促進し、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、以下の行動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、レッドチーミング等の評価方法を組み合わせて、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に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。これらの対策は文書化され、定期的に更新される技術文書によってサポートされるべきである。

このようなテストは、リスクと脆弱性を特定するため、また、偶発的か意図的かを問わず、セキュリティ、安全性、社会的リスク、その他のリスクに対処するための行動を通知するために、安全な環境で実施されるべきであり、また、特に導入前及び市場投入前等のAIライフサイクル全体におけるいくつかのチェックポイントで実施されるべきである。テスト措置の設計と実施において、組織は以下のリスクに適切に注意を払うことを約束する：

- > 高度なAIシステムが、非国家主体も含め、兵器の開発、設計の取得、使用への参入障壁を低くする方法等、化学、生物、放射性、核のリスク。
- > 攻撃的サイバー能力とは、システムが脆弱性の発見、悪用、又は作戦上の利用を可能にす

る方法等であり、そのような能力の有用な防御的応用の可能性があり、システムに含めることが適切であるかもしれないことを念頭に置くこと。

- >健康及び/又は安全に対するリスク。システムの相互作用やツールの使用による影響を含み、例えば物理的なシステムを制御したり、重要なインフラに干渉したりする能力を含む。
- > モデルが自分自身のコピーを作ったり、“自己複製”したり、他のモデルを訓練したりすることによるリスク。
- > 高度な AI システムやモデルが有害な偏見や差別を生じさせたり、プライバシーやデータ保護等適用される法的枠組みへの違反につながったりする可能性等、社会的リスクや個人やコミュニティに対するリスク。
- > 偽情報の助長やプライバシーの侵害等、民主主義の価値や人権に対する脅威。
- > 特定の事象が連鎖反応を引き起こし、都市全体、領域活動全体、地域社会全体にまで重大な悪影響を及ぼすリスク。

各組織は、セクターを超えた関係者と協力して、これらのリスク、特にシステミック・リスクに対処するための緩和策を評価し、採用することを約束する。

また、これらのコミットメントに取り組む組織は、高度な AI システムのセキュリティ、安全性、偏見と偽情報、公平性、説明可能性と解釈可能性、透明性に関する研究と投資を促進し、悪用に対する先進的 AI システムの堅牢性と信頼性を高めることに努めるべきである。

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、責任を持って弱点を開示するインセンティブを与えるための報奨金制度、コンテスト、賞品等を通じて、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、アカウント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。

これらの報告書、使用説明書、及び関連する技術的文書は、適宜、最新に保たれるべきであり、例えば、以下のようなものが含まれるべきである；

- > 潜在的な安全性、セキュリティ、社会的リスク及び人権に対するリスクについて実施された評価の詳細。
- > 適切な使用領域に影響を及ぼすモデル／システムの能力と性能上の重大な限界。
- > 有害な偏見、差別、プライバシーや個人情報保護への脅威、公平性への影響等、モデルやシステムが安全性や社会に及ぼす影響やリスクについての議論と評価。
- > 開発段階以降のモデル／システムの適合性を評価するために実施されたレッドチーミングの結果。

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使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、技術文書や使用説明書等の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。

各組織は、高度な AI システムの安全性、セキュリティ、信頼性を確保するための共有の基準、ツール、メカニズム、ベストプラクティスを開発し、推進し、必要に応じて採用するためのメカニズムを構築するか、それに参加すべきである。

これには、特に安全性と社会に重大なリスクをもたらす高度な AI システムに関して、AI のライフサイクル全体にわたって適切な文書化と透明性を確保することも含まれるべきである。

組織は、高度な AI システムの安全性、セキュリティ、信頼性を向上させる観点から、AI のライフサイクル全体にわたって他の組織と協力し、関連情報を共有し、社会に報告すべきである。また、組織は、必要に応じて、関連する公的機関とも連携し、前述の情報を共有すべきである。

このような報告は、知的財産権を保護すべきである。

5 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

組織は、AI のライフサイクルを通じて、実現可能であれば、リスクを特定し、評価し、予防し、対処するための説明責任とガバナンスのプロセス等を含む、リスク管理とガバナンスの方針を策定し、開示し、実施するための適切な組織的メカニズムを導入すべきである。

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、これらの方針を実施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。

リスク管理方針は、リスクベースのアプローチに従って策定されるべきであり、AI システムに関連する様々なリスクに対処するために、適切かつ関連する AI のライフサイクル全体にわたってリスク管理の枠組みを適用すべきであり、また、方針は定期的に更新されるべきである。

組織は、職員が自らの責任及び組織のリスク管理慣行を熟知していることを確保するための方針、手順及び研修を確立すべきである。

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。

また、高度な AI システムのサイバーセキュリティが関連する環境及び関連するリスクに照らして適切であることを確保するため、サイバーセキュリティリスクの評価を実施し、サイバーセキュリティポリシー及び適切な技術的・制度的解決策を実施することも含まれる。また、組織は、高度 AI システムのモデルの重みの保管と作業を、アクセスが制限された適切で安全な環境で行うことを

義務付け、無許可で公開されるリスクと不正アクセスされるリスクの両方を低減するための対策を講じる必要がある。これには、脆弱性管理プロセスを導入し、セキュリティ対策を定期的に見直して、それらが高い水準に維持され、リスクに対処するのに適切であり続けることを保証するコミットメントが含まれる。

これにはさらに、例えば、非公開のモデルの重みへのアクセスの制限等の、最も貴重な知的財産や企業秘密に対する保護と整合性のある、強固な内部脅威検知プログラムの確立も含まれる。

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、この分野の状況を前進させるために、必要に応じて、協力し、研究に投資すべきである。

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。

組織は、AI の安全性、セキュリティ、信頼性の向上を支援し、民主的価値の維持、人権の尊重、子どもや社会的弱者の保護、知的財産権とプライバシーの保護、有害な偏見、偽・誤情報、情報操作の回避等の重要なリスクに対処する優先的な研究を実施し、協力し、投資することにコミットする。組織はまた、適切な緩和ツールの開発に投資することにコミットし、環境や気候への影響を含む高度な AI システムのリスクを積極的に管理し、その便益が実現されるよう努力する。

組織は、リスク緩和に関する研究とベストプラクティスを共有することが奨励される。

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に 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、高度な AI システムの利用から利益を得ることができるようにし、個人や地域社会がこれらの技術の性質、能力、限界、影響をよりよく理解できるようにするデジタル・リテラシーのイニシアティブを支援するため、学生や労働者を含む一般市民の教育と訓練を促進すべきである。組織は、市民社会やコミュニティ・グループと協力して、優先課題を特定し、世界最大の課題に取り組むための革新的な解決策を開発すべきである。

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。

組織は、組織のテスト方法、コンテンツの認証及び来歴メカニズム、サイバーセキュリティポリシー、公開報告、その他の手段を開発する際にも、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)と協力することが奨励される。特に、AI が生成したコンテンツと AI 以外が生成したコンテンツを利用者が区別できるようにするための、相互運用可能な国際的な技術標準や枠組みの開発に取り組むことが奨励される。

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、有害な偏見を軽減するために、訓練データやデータ収集等、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。

適切な対策には、透明性、プライバシーを保護するトレーニング技術、及び／又はシステムが機密データや機微データを漏らさないようにするためのテストとファインチューニングが含まれる。

組織は、著作権で保護されたコンテンツを含め、プライバシーや知的財産に関する権利を尊重するために、適切なセーフガードを導入することが奨励される。

組織はまた、適用される法的枠組みを遵守すべきである。



Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems

On the basis of the Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI systems, the Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems aims to promote safe, secure, and trustworthy AI worldwide and will provide voluntary guidance for actions by organizations developing the most advanced AI systems, including the most advanced foundation models and generative AI systems (henceforth "advanced AI systems").

Organizations should follow these actions in line with a risk-based approach.

Organizations that may endorse this Code of Conduct may include, among others, entities from academia, civil society, the private sector, and/or the public sector.

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

This document will be reviewed and updated as necessary, including through ongoing inclusive multistakeholder consultations, in order to ensure it remains fit for purpose and responsive to this rapidly evolving technology.

Different jurisdictions may take their own unique approaches to implementing these actions 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.



In addition, we encourage organizations to set up internal AI governance structures and policies, including self-assessment mechanisms, to facilitate a responsible and accountable approach to implementation of these actions and in AI development.

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 ways that undermine democratic values, are particularly harmful to individuals or communities, facilitate terrorism, promote 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 ensure 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 actions, in a manner that is 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 for evaluations, 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. These measures should be documented and supported by regularly updated technical documentation.

This testing should take place in secure environments and be performed at several checkpoints throughout the AI lifecycle in particular before deployment and placement on the market to identify risks and vulnerabilities, and to inform action to address the identified AI risks to security, safety and societal and other risks, whether accidental or intentional. In designing and implementing testing

measures, organizations commit to devote attention to the following risks as appropriate:

- > Chemical, biological, radiological, and nuclear risks, such as the ways in which advanced AI systems can lower barriers to entry, including for non-state actors, for weapons development, design acquisition, or use.
- > Offensive cyber capabilities, such as the ways in which systems can enable vulnerability discovery, exploitation, or operational use, bearing in mind that such capabilities could also have useful defensive applications and might be appropriate to include in a system.
- > Risks to health and/or Safety, including the effects of system interaction and tool use, including for example the capacity to control physical systems and interfere with critical infrastructure.
- > Risks from models of making copies of themselves or “self-replicating” or training other models.
- > Societal risks, as well as risks to individuals and communities such as the ways in which advanced AI systems or models can give rise to harmful bias and discrimination or lead to violation of applicable legal frameworks, including on privacy and data protection.
- > Threats to democratic values and human rights, including the facilitation of disinformation or harming privacy.
- > Risk that a particular event could lead to a chain reaction with considerable negative effects that could affect up to an entire city, an entire domain activity or an entire community.

Organizations commit to work in collaboration with relevant actors across sectors, to assess and adopt mitigation measures to address these risks, in particular systemic risks.

Organizations making these commitments should also endeavor to advance research and investment on the security, safety, bias and disinformation, fairness, explainability and interpretability, and transparency of advanced AI systems and on increasing robustness and trustworthiness of advanced AI systems against misuse.

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 such as through bounty systems, contests, or prizes to incentivize the responsible disclosure of weaknesses. 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.

These reports, instruction for use and relevant technical documentation, as appropriate as, should be kept up-to-date and should include, for example;

- > Details of the evaluations conducted for potential safety, security, and societal risks, as well as risks to human rights,
- > Capacities of a model/system and significant limitations in performance that have implications for the domains of appropriate use,
- > Discussion and assessment of the model's or system's effects and risks to safety and society such as harmful bias, discrimination, threats to protection of privacy or personal data, and effects on fairness, and
- > The results of red-teaming conducted to evaluate the model's/system's fitness for moving beyond the development stage.

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 such as technical documentation and instructions for use.

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.

Organizations should establish or join mechanisms to develop, advance, and adopt, where appropriate, shared standards, tools, mechanisms, and best practices for ensuring the safety, security, and trustworthiness of advanced AI systems.

This should also include ensuring appropriate and relevant documentation and transparency across the AI lifecycle in particular for advanced AI systems that cause significant risks to safety and society.

Organizations should collaborate with other organizations across the AI lifecycle to share and report relevant information to the public with a view to advancing safety, security and trustworthiness of advanced AI systems. Organizations should also collaborate and share the aforementioned information with relevant public authorities, as appropriate.

Such reporting should safeguard intellectual property rights.

5 Develop, implement and disclose AI governance and risk management policies, grounded in a risk-based approach – including privacy policies, and mitigation measures.

Organizations should put in place appropriate organizational mechanisms to develop, disclose and implement risk management and governance policies, including for example accountability and governance processes to identify, assess, prevent, and address risks, where feasible throughout the AI lifecycle.

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.

The risk management policies should be developed in accordance with a risk-based approach and apply a risk management framework across the AI lifecycle as appropriate and relevant, to address the range of risks associated with AI systems, and policies should also be regularly updated.

Organizations should establish policies, procedures, and training to ensure that staff are familiar with their duties and the organization's risk management practices

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.

This also includes performing an assessment of cybersecurity risks and implementing cybersecurity policies and adequate technical and institutional solutions to ensure that the cybersecurity of advanced AI systems is appropriate to the relevant circumstances and the risks involved. Organizations should also have in place measures to require storing and working with the model weights of advanced AI systems in an appropriately secure environment with limited access to reduce both the risk of unsanctioned release and the risk of unauthorized access. This includes a commitment to have in place a vulnerability management process and to regularly review security measures to ensure they are maintained to a high standard and remain suitable to address risks.

This further includes establishing a robust insider threat detection program consistent with protections provided for their most valuable intellectual property and trade secrets, for example, by limiting access to proprietary and unreleased model weights.

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 and 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 should collaborate and invest in research, as appropriate, to advance the state of the field.

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.

Organizations commit to conducting, collaborating on and investing in research that supports the advancement of AI safety, security, trustworthiness and addressing key risks, such as prioritizing research on upholding democratic values, respecting human rights, protecting children and vulnerable groups, safeguarding intellectual property rights and privacy, and avoiding harmful bias, mis- and disinformation, and information manipulation. Organizations also commit to invest in developing appropriate mitigation tools, and work to proactively manage the risks of advanced AI systems, including environmental and climate impacts, so that their benefits can be realized.

Organizations are encouraged to share research and best practices on risk mitigation.

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 that promote the education and training of the public, including students and workers, to enable them to benefit from the use of advanced AI systems, and to help individuals and communities better understand the nature, capabilities, limitations, and impact of these technologies. Organizations should work with civil society and community groups to identify priority challenges and develop innovative solutions to address the world's greatest challenges.

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

Organizations are encouraged to contribute to the development and, where appropriate, use of international technical standards and best practices, including for watermarking, and working with

Standards Development Organizations (SDOs), also when developing organizations' testing methodologies, content authentication and provenance mechanisms, cybersecurity policies, public reporting, and other measures. In particular, organizations also are encouraged to work to develop interoperable international technical standards and frameworks to help users distinguish content generated by AI from non-AI generated content.

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 measures could include transparency, privacy-preserving training techniques, and/or testing and fine-tuning to ensure that systems do not divulge confidential or sensitive data.

Organizations are encouraged to implement appropriate safeguards, to respect rights related to privacy and intellectual property, including copyright-protected content.

Organizations should also comply with applicable legal frameworks.