（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、学術界、市民社会、民間部門、公共部門の主体等が含まれうる。

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。
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。