安全、革新的で包摂的なAIの発展のためのソウル閣僚声明
（2024年5月22日のAIソウルサミット）

前文

1. 我々、オーストラリア、カナダ、チリ、フランス、ドイツ、インド、インドネシア、イスラエル、イタリア、日本、ケニア、メキシコ、オランダ、ナイジェリア、ニュージーランド、フィリピン、大韓民国、ルワンダ、サウジアラビア王国、シンガポール共和国、スペイン、スイス、トルコ、ウクライナ、アラブ首長国連邦、英国、米国の閣僚、及び欧州連合代表は、2024 年 5 月 22 日に開催された AI ソウルサミットの閣僚セッションにおける、安全、安心で信頼できる AI 及び持続可能な AI 開発を促進するための AI ガバナンスのアプローチに関する議論に基づき、AI 技術の急速な進展及び我々の社会・経済への影響に対応するための国際的な協調的アプローチの必要性をここに確認する。

2. 我々は、ブレッチリー・パークで開催された第 1 回 AI 安全性サミットの成果を高く評価する。AI ソウルサミットの 1 日目に議論された、安全性、革新性、包摂性という相互に関連する 3 つの優先事項に基づいて、我々は今、それらを実施するための実行可能な項目に焦点を当てようとしている。我々は、それぞれの国内における能力の違いを認識し、AI のリスクに対処しつつ、全ての人々にとっての AI の恩恵を引き出すために有意義な措置を講じるという共通の意思を確認する。

安全性

1.1 現在及び最先端の AI モデルやシステムの導入・利用によってもたらされるリスク並びに今後設計・開発・導入・利用される可能性のあるものによりもたされるリスクを含め、AI のあらゆるリスクを予防することが不可欠である。AI の安全性及び安心のための原則には、透明性、解釈可能性、説明可能性、プライバシーと説明責任、人間による意味のある監視、効果的なデータ管理と保護が含まれる。私たちは、現在及び最先端の AI を開発・導入している組織を含む全ての関係主体が、AI ライフサイクル全体を通じて、生じ得る悪影響を評価、予防、緩和、修正するよう努めることで、説明責任及び透明性を促進することを奨励する。さらに、我々は、AI が安全、安心で信頼できる方法で、全ての人の利益のために、及び適用される国内の及び国際的な枠組みに沿った形で、設計、開発、導入、利用されるような環境を促進することを全ての関係主体に対して奨励する。

1.2 我々は、それぞれの管轄区域において、商業的又は公的に利用可能な最先端 AI モデル又はシステムの設計、開発、導入及び利用によってもたらされるリスクを管理するための枠組みを確立する我々の役割を認識する。我々は、我々の管轄区域において開発された最先端 AI モデル又はシステムが重大なリスクをもたらす可能性がある場合、そのモデル又はシステムに対する信頼できる外部評価を奨励する我々の増大する役割を認識する。我々はさらに、最先端 AI モデル又はシステムの設計、開発、導入及び利用によってもたらされるリスクが、適切な軽減なしには重大なものとなる閾値を特定する上で、民間セクター、市民社会、学術界及び国際社会とのパートナーシップ における我々の役割を認識する。最先端 AI モデル又はシステムがもたらすリスクを評価するための基準には、能力、限界及び傾向、悪意のある敵対的な攻撃及び操作に対する頑健性を含む実装されたセーフガード、予見可能な利用及び誤用、AI モデルが統合される可能性のある広範なシステムを含む導入の状況、影響範囲、その他の関連するリスク要因の考慮が含まれる。

1.3 最先端 AI モデル又はシステムの設計、開発、導入及び利用がもたらすリスクの評価には、適切な緩和策を講じない限り、重大なリスクをもたらす可能性のあるモデル又はシステムの能力をその状況において定義し、測定することが含まれる場合がある。我々は、そのような重大なリスクは、化学・生物兵器及びその運搬手段の開発、生産、取得、使用を進める上で、非国家主体を有意に支援するモデル及びシステムの潜在的能力によってもたらされる可能性があることを認識する。我々は、各国の義務に従って、化学兵器禁止条約及び生物兵器禁止条約、国連安全保障理事会決議第 1540 号、国際人権法等の関連する国際法に従って行動することの継続的な重要性を確認する。我々は、AI の安全性と安心を促進するための多国間協議の重要性を強調する。

1.4 我々は、さらに、セーフガードの回避、操作及び欺瞞、又は人間の明示的な承認若しくは許可なく行われる自律的な複製及び適応を含む、人間による監視を回避する潜在的なモデル又はシステムの能力若しくは傾向によって、このような重大なリスクがもたらされる可能性があることを認識する。我々は、高度に進化したエージェント能力を有する最先端 AI モデル又はシステムによるリスクに関して、更なる実証的データを収集することの重要性に留意すると同時に、人間による意味のある監視等の適切なセーフガードを実施するために、最先端 AI を開発・導入する組織と協力することを含め、そのようなモデル又はシステムの誤用又はずれを防止することの必要性を認める。

1.5 我々は、最先端 AI モデルやシステムのリスクに対処するための開発者との建設的な対話の重要性を認識し、これらのシステムの安全性に対する開発者の特別な責任を再確認する。我々はさらに、AI ライフサイクル全体を通じて安全性とセキュリティを考慮することが急務であることを認識する。

1.6 我々は、AI のリスク管理に関する国際協力を強化し、AI の安全性・セキュリティの領域における世界的な理解を深めるために、AI セーフティ・インスティテュートやその他の関連機関が独自の役割を果たすことを確認する。我々は、AI セーフティ・インスティテュート又はその他の関連機関を通じて、必要に応じてベストプラクティス及び評価データセットを共有し、安全性試験ガイドラインの策定において協力する予定である。我々は、AI セーフティ・インスティテュートやその他の関連機関とのパートナーシップを構築することを含め、AI 安全性活動の相互運用可能性を目指すと同時に、世界中の文化的・言語的多様性を考慮した試験方法の必要性を認識する。

革新性

2.1 我々は、AI が我々の経済と社会にもたらす潜在的な利益を最大化することを目的として、イノベーションと AI 産業のエコシステムの開発を促進するガバナンス・アプローチの重要性を認識する。さらに、政府の役割は、AI の革新のための財政投資、研究開発、人材開発を優先させるだけでなく、AI の安全、安心で信頼できる開発と導入のために、個人情報、著作権 その他の知的財産 の保護を含む法的・制度的枠組みを含むガバナンス枠組を検討することであると認識する。

2.2 我々は、行政、福祉、教育、医療などの分野を含む公共部門にとって、AI が変革的な利益をもたらすことを認識している。これらの利益には、利用しやすいデジタル・サービスや、公共サービスを利用する際の市民の体験を向上させる自動化された手続を通じて、AI を効率的かつ効果的に利用することが含まれる。さらに、我々は、生産性を変革し、権利と安全を守りながら従業員の負担を軽減し、価値創造の新たな道を切り開くために、製造、物流、金融などの主要産業分野における AI の導入を支援する。

2.3 我々は、特に、知的財産権を尊重し保護しつつ、特に中小企業、スタートアップ、学術界、大学、さらには個人の AI 関連リソースへのアクセスを容易にすることにより、AI 主導のイノベーションを促す環境を支援することにコミットしている。また、研究者がそれぞれの研究分野で AI を活用できるようにするため、また、個人の創造的な努力を豊かにするツールとして AI の責任ある活用を促進するため、AI 関連リソースの利用可能性を高めることにコミットしている。

2.4 我々は、AI イノベーションのエコシステムにおける持続可能性とレジリエンスの重要性を認識する。この観点から、我々は、AI の開発者及び導入者が、エネルギーや資源の消費といった潜在的な環境フットプリントを考慮することを奨励する。我々は、イノベーションと生産性を向上させるために、AI を自信を持って使用し開発できるように、労働力をどのようにスキルアップ及び再スキルアップさせることができるかについての方策を探るための協力的な取り組みを歓迎する。さらに、我々は、AI 開発やサービスを通じて、低消費電力の AI チップの適用や環境に配慮したデータセンターの運用など、資源効率の高い AI モデルやシステム、インプットの開発・利用を促進する企業の取り組みを奨励する。

包摂性

3.1 包摂的なデジタルトランスフォーメーションを促進する取組において、我々は、AI の恩恵は公平に共有されるべきであると認識する。我々は、社会的弱者を含む全ての人々が AI の恩恵を活用できるよう、共通のビジョンを推進する。我々は、共通の価値観と相互信頼を育むため、AI システムの包摂的な開発と、安全、安心で信頼できる AI 技術の活用を促進するために協力する。我々は、特に、人権及び基本的自由の保護、社会的セーフティネットの強化、並びに災害及び事故を含む様々なリスクからの安全の確保において、全ての人の利益となる AI の可能性を認識する。

3.2 包摂性という共通の目標を推進するため、我々は、AI システムに関するキャパシティビルディングやデジタル・リテラシーの向上を通じた AI 教育の推進にコミットし、国家間及び国内の AI 及びデジタル・デバイドの解消に貢献する。我々は、AI の設計、開発、利用における能力を高めるために、発展途上国を含む共同研究や人材開発における国際協力を強化する必要性を認識している。我々は、設計、開発、導入、利用といった AI ライフサイクルにおいて、社会文化的、言語的多様性が反映され、促進されることを確保することを目指す。

3.3 我々は、気候変動、グローバルヘルス、食料・エネルギー安全保障、教育など、世界最大の課題を解決するための重要な進歩をもたらす可能性を認識し、AI 技術の進歩を支援・促進することにコミットする。我々は、さらに、持続可能な開発目標の達成に向けた進捗を加速し、世界的な共通の利益と発展を促進することを目的とした共同の努力と議論への途上国の参加を奨励することにより、包摂的なガバナンス・アプローチを促進することを目指す。

今後の方向性

1. 我々は、AI ソウルサミットの共同議長国である韓国及び英国が、安全、持続可能性及びレジリエンスの課題に関して行った努力を称賛する。

2. 我々は、高度なAIの安全性に関する独立した国際科学中間報告書の公表と、最先端AIに関連するリスクに関する証拠に基づく理解の共有を促進するための作業に留意する。我々は、AIのリスクに関する将来の証拠に基づく報告書を前進させるために協力することを決意し、フランスで開催されるAIアクション・サミットに先駆けて、高度なAIの安全性に関する国際科学報告書の最終公表を期待する。

3. 我々は、AI 安全性に関する科学を進展させ、特定のリスクに関してより多くの実証的データを収集する必要性を認めると同時に、重大なリスクをもたらす可能性のある能力に関して、我々の集合的理解を経験に基づく事前措置に反映させる必要性を認識する。また、フランスで開催される AI アクション・サミットに先駆けて検討のための提案を作成するという願いとともに、適切な緩和策を講じなければ、最先端 AI モデルやシステムの設計、開発、導入、利用によってもたらされるリスクのレベルが重大なものとなる閾値を特定し、また、重大なリスクをもたらす可能性のある最先端 AI モデルやシステムの能力を定義するため、民間セクター、市民社会、学術界と協力する。

4. 我々は、AI の設計、開発、導入、利用を、その利点を善に活用する形で導くという共通の意図を再確認する。安全性、革新性、包摂性が相互に関連する目標であることを認識した上で、我々は、AI ガバナンスに関する議論を進め、万人の利益のために安全、安心で信頼できる AI を促進するための協力を継続することを期待する。