汎用AIの初稿
行動規範

議長および副議長による開会の挨拶

4つのワーキンググループの議長および副議長として、私たちはここに、AI法に基づく汎用AI実践規範（以下「規範」）の初版を発表します。実践規範の全体会議参加者およびオブザーバーからの書面によるフィードバックは、専用プラットフォーム（Futurium）のフォームを通じて、11月28日（木）12:00 CETまでに受け付けています。

この規範の最初の草案では、緊密に連携して作業する 4 つのワーキング グループを通じて、汎用 AI モデルの提供者とシステムリスクを伴う汎用 AI モデルの提供者にとって重要な考慮事項を取り上げています。

- ワーキンググループ1: 透明性と著作権関連ルール
- ワーキンググループ2: システムリスクのリスク特定と評価
- ワーキンググループ3: システミックリスクの技術的リスク軽減
- ワーキンググループ4: システミックリスクに対するガバナンスリスク軽減

この最初の草案は、さらなる改良のための基礎として提示します。ワーキング グループ内での反復的な内部討論プロセスと、利害関係者からの追加の外部インプットを経て、対策が追加、削除、または変更される可能性があります。利害関係者には、この文書をレビューしてフィードバックを提供し、汎用 AI モデルの開発と展開の将来を導く上で重要な役割を果たすコードの最終版の作成にご協力いただくようお願いします。

規範の指針となる原則と目的を概説した高レベルの起草計画も含まれています。
最初の草案は詳細が薄いですが、このアプローチは、特定のサブメジャーと主要業績指標（KPI）に関する徹底的な審議を継続しながら、最終的な規範の潜在的な形式と内容の方向性を利害関係者に明確に伝えることを目的としています。審議をさらに詳しく知るために、将来の草案で進歩を目指すいくつかの領域を強調する未解決の質問を追加しました。これは、フィードバックと提案を導く目的にも役立ちます。

さまざまな利害関係者が引き続き効果的に参加できるようにします。

EU AI法は2024年8月1日に発効し、コードの最終版は2025年5月1日までに準備する必要があるとされています。ここで提示された最初の草案は、2025年以降に開発およびリリースされる次世代モデルにも適した「将来を見据えた」コードを提供することを目指しています。2024年10月に始まった私たちの作業には、さまざまな関係者からの意見を統合し、反復的な議論を行うことが含まれていました。

この最初の草案を作成するにあたり、議長および副議長は、本規範の範囲内の事項については、主に AI 法の規定に従ってきました。したがって、本規範に含まれる文脈および定義が別段の定めをしない限り、本規範で使用されている用語は AI 法の同一の用語を参照しています。これには、AI 法第 56 条 (1) の、さまざまな既存の国際的アプローチを考慮に入れるようにという指示が含まれます。この最初の草案では AI 法の規定を網羅的に参照していませんが、今後の反復でそうする予定です。

この規範の初稿は、業界、学界、市民社会からの何百人もの参加者による共同作業の結果です。また、AI ガバナンス、国際的なアプローチ、EU 法の実践規範 (偽情報に関する実践規範など) に関する最新の文献、およびワーキング グループ メンバーの専門知識と経験も参考にしています。

開発プロセスの主な特徴は次のとおりです。

- これまでに430件近くの意見が寄せられた複数の利害関係者による協議
- 専門知識、経験、独立性、地理的および性別の多様性を確保するために選ばれた議長と副議長が率いる4つの専門作業部会
- 2024年10月から2025年4月にかけて議論と草案作成セッションを開催

現在の草案を洗練し、改善するには、外部および内部の両方で協議と審議を行うための追加の時間が必要になります。独立した議長と副議長のグループとして、私たちはこのプロセスを可能な限り透明かつ利害関係者がアクセスしやすいものにするよう努め、作業グループ内で主要な問題を調整および議論する十分な時間を取りながら、できるだけ早く作業と考え方を共有することを目指しています。私たちは、皆さんの継続的な積極的な協力と建設的な批判を期待しています。

行動規範の全体会議参加者およびオブザーバーからの書面によるフィードバックを、専用プラットフォーム (Futurium) のフォームを通じて、11 月 28 日木曜日 12:00 CET までにお待ちしています。

ご支援ありがとうございます！

ヌリア•オリバー ワーキンググループ1 共同議長	アレクサンダー•ペウケルト ワーキンググループ1 共同議長	マティアス•サムワルド ワーキンググループ2 椅子	ジョシュア•ベンジオ ワーキンググループ3 椅子	マリエッチェ•シャアケ ワーキンググループ4 椅子
リシ•ボンマサニ ワーキンググループ1 副議長	セリーヌ•カステ=ルナール ワーキンググループ1 副議長	マルタ•ジオシ ワーキンググループ2 副議長	ダニエル•プリビテラ ワーキンググループ3 副議長	リューエルとも ワーキンググループ4 副議長
		アレクサンダー•ザッヘル ワーキンググループ2 副議長	ニタルシャン•ラージクマール ワーキンググループ3 副議長	マルクス•アンデルユング ワーキンググループ4 副議長

計画と原則の立案

現段階では、この最初の草案には、最終的に採択される規範に含まれる粒度レベルは含まれていません。その理由は、i) 我々は規範の構造と原則について幅広い合意を目指していること、ii) この最初の草案で必要とされるレベルの検討を伴う詳細な提案を作成する時間が十分になかったこと、iii) 我々は、継続的に最新の動向を反映するために、最後に (草案の) 規範の詳細を更新するためです。この最初の草案は全体的にハイレベルな内容であるにもかかわらず、我々の方向性を示すために、いくつかのコミットメントの下に、より具体的な規定を垣間見させ、規範の将来の草案で同様の規定がどのような形式と詳細をとる可能性があるかを示しています。この規範におけるコミットメントの構造は、対策、サブ対策、KPI の降順の階層構造になっています。これらのいずれか、特に KPI が欠落している場合、これは最終決定ではなく、時間的制約とこの最初の草案のハイレベルな内容の結果です。さらに、この最初の草案には、規範がどのようにレビューされ、更新されるかについてのセクションがまだ含まれていません。これは、規範草案の後の反復に組み込まれる予定です。

また、本規範を起草する際に従うことを提案するいくつかの高レベルの原則を以下に示しています。

I. EU の原則と価値観との整合 –対策、サブ対策、KPI は、欧州連合基本権憲章、欧州連合条約、欧州連合機能条約を含む EU 法に定められた、EU の一般原則と価値観に沿ったものでなければなりません。

II. AI法と国際的アプローチとの整合 –措置、サブ措置、KPIはAI法の適切な適用に寄与するものであるべきである。これには以下の事項を考慮することが含まれる。
AI法第56条第1項に従って、国際的なアプローチ（AI安全研究所または標準化団体が策定した標準または指標を含む）。

III. リスクへの比例性 –指標、サブ指標、KPIはリスクに比例するべきである。
つまり、(a) 望ましい目的を達成するのに適切であり、(b) 望ましい目的を達成するために必要であり、(c) 達成しようとする目的に比べて過度の負担を課さないものでなければならない。比例性の具体的な適用例には、次のようなものがある。

a) より重大なリスクや重大な危害の不確実なリスクに対しては、対策、サブ対策、KPI をより厳格にすべきです。この規範は、例えば、重大なリスクに関連するサブ対策ごとに多数の KPI を提案することでこれを実現し、汎用 AI モデルの提供者に、その重大なリスクを軽減するための措置を講じるか、重大なリスクが発生する可能性が極めて低いことを確実に実証するよう要求します。この規範は、

また、「if-then」要件を使用するなど、リスク軽減サブ対策をリスク評価 KPI に結び付けることもできます。たとえば、システミックリスクのある汎用 AI モデルが機能 X を備えていると評価された場合、Z KPI に基づいて Y リスク軽減策を実施する必要があります。

b) サブ対策と KPI は具体的である必要があります。対策はサブ対策やサブ対策への準拠を証明する KPI よりも一般性が高いレベルで記述される可能性があることを認めます。ただし、汎用 AI モデル プロバイダーは、KPI によって証明されるサブ対策を適宜満たす方法を可能な限り明確に理解する必要があります。サブ対策と KPI は、回避や誤った指定に対しても堅牢である必要があります。

この規範は、例えば、代理権の不必要な使用を避けることによってこれを実現することができる。

用語や指標。AI オフィスは、回避やその他の誤った指定の影響を受ける可能性のあるサブ指標と KPI を監視および確認します。

c) 対策、サブ対策、KPI は、該当する場合、リスクの種類、配布戦略、展開コンテキスト、およびリスク レベルに影響を及ぼす可能性のあるその他の要因、およびリスクをどのように評価して軽減する必要があるかを区別する必要があります。たとえば、システミック リスクの評価と軽減に関連する対策、サブ対策、KPI では、意図的なリスクと意図的でないリスク (不整合を含む) を区別する必要がある場合があり、一部の種類のリスク、配布戦略 (例: オープンソース)、展開コンテキストに対しては、他のものよりも具体的または厳格になる場合があります。

IV.将来性– サブメジャーと KPI は、優れた情報に基づいてコンプライアンスの評価を改善する AI オフィスの能力を維持する必要があります。さらに、サブメジャーと KPI を更新するプロセスでは、急速な技術変化により、機敏な規制の開発と変更が必要になる可能性があることを想定する必要があります。したがって、具体的な要件と、技術と業界の発展に合わせてルールを適応および更新する柔軟性との間でバランスを取る必要があります。規範は、たとえば、プロバイダーが監視して自分で検討することが期待される動的な情報源を参照することでこれを実現できます。このような情報源の例には、インシデント データベース、コンセンサス標準、リスク レジスタ、リスク管理フレームワーク、AI オフィスのガイダンスなどがあります。規範は、サブメジャーと KPI を通常よりも迅速に更新できるように努めます。エージェント AI システムで使用されるモデルなど、新しいサブメジャーと KPI のセットを必要とするモデルの種類を明確にする必要もあります。

V.汎用AIモデルプロバイダーの規模への比例性 –汎用AIモデルプロバイダーに適用される義務に関連する措置とKPIは、汎用AIモデルプロバイダーの規模を適切に考慮し、必要に応じて、AI開発の最前線にいる企業よりも資金力の少ない中小企業や新興企業向けに簡素化されたコンプライアンス方法を許可する必要があります。システミックリスクのある汎用AIモデルプロバイダーに適用される義務に関連するKPIは、必要に応じて、プロバイダーの規模と能力の違いも反映する必要があります。

VI. AI安全エコシステムのサポートと成長 – AIの安全性エコシステムの開発、採用、
汎用AIモデルの安全性とガバナンスは世界的な課題です。この草案の多くの措置は、モデルプロバイダー間で汎用AIの安全性インフラストラクチャとベストプラクティスを共有するだけでなく、市民社会、学術界、第三者、政府機関の貢献をさらに可能にするなど、さまざまな利害関係者間の協力を可能にし、サポートすることを目的としています。このため、私たちは利害関係者間の透明性をさらに高め、知識を共有し、AI法第56条(1)(3)および第116条の序文に沿って、AIの安全性に関する集合的で堅牢な証拠ベースの構築に協力する努力を強化することを奨励しています。また、オープンソースモデルがAIの安全性エコシステムの発展に与えたプラスの影響も認識しています。

目次

議長および副議長による開会の辞 1
計画立案と原則..... 3
目次..... 5
I. 序文..... 6
II. 汎用AIモデルの提供者に対する規則..... 8
透明性..... 10
対策1. AIオフィスの文書化..... 10
対策2. 下流プロバイダー向けの文書化..... 10
付録: 許容使用ポリシーの必須要素 13
著作権に関する規則..... 14
対策3. 著作権ポリシーを導入する 14
対策4. TDM例外の制限の遵守.....15
対策5. 透明性 16
III. システミックリスクの分類 17
対策6. 分類法..... 17
IV. システムリスクを伴う汎用AIモデルの提供者に対する規則 20
対策7. 安全とセキュリティの枠組み..... 21
システムリスクを伴う汎用 AI モデルプロバイダーのリスク評価 ....
対策8. リスクの特定 22
対策9. リスク分析..... 22
措置10. 証拠収集..... 23
対策11. リスク評価ライフサイクル 25
システムリスクを伴う汎用AIモデルプロバイダー向けの技術的リスク軽減
.... 27
対策12. 緩和策..... 27
対策13. 安全とセキュリティに関する報告書..... 28
対策14. 開発と展開の決定..... 29
システムリスクを伴う汎用 AI モデルプロバイダー向けのガバナンスリスク軽
減 30
対策15. システム的リスクの所有権..... 30
対策16. 遵守と適切性の評価..... 30
対策17. 独立した専門家による体系的リスクと緩和の評価..... 31
対策18. 重大インシデントの報告..... 32
措置19. 内部告発の保護..... 33
措置20. 通知..... 33
対策21. 文書化 34
措置22. 公的透明性..... 35
結論..... 36

I. 序文

一方：

a) この行動規範（コード）の署名者は、EUにおけるAIの有害な影響から、健康、安全、憲章に定められた基本的権利（民主主義、法の支配、環境保護を含む）を高いレベルで保護し、イノベーションを支援しながら、域内市場の機能改善、人間中心で信頼できる人工知能（AI）の規制のための公平な競争の場の創出の重要性を認識している。

AI法（法）第1条第1項に強調されているとおり、本規約はこのような文脈で解釈されるものとする。

b) 本規範は、汎用AIモデルおよびシステムリスクを伴う汎用AIモデルの提供者に対して、AI法第53条および第55条に定められた義務の遵守に関するガイダンスを提供します。

c) 本規約が汎用AIモデルの提供者に言及する場合は、システムリスクを伴う汎用AIモデルの提供者も対象とします。本規約が汎用AIモデルの提供者に言及する場合は、システムリスクを伴う汎用AIモデルの提供者も対象とします。
システムリスクのある汎用 AI モデルについては、他の汎用 AI モデルの提供者は対象としません。

d) 署名者は、本規範が、システムリスクを伴うGPAIモデルおよび汎用AIモデルのプロバイダーにとって、AI法の遵守を証明するためのガイド文書として機能することを認識する一方で、本規範の遵守がAI法の遵守の決定的な証拠を構成するものではないことを認識しています。

e) 署名者は、AI 事務局および理事会による規範の妥当性の定期的な監視および評価を容易にするために、規範の実施状況および結果を報告することの重要性を認識しています。

f) 本規範は、AI オフィスによる定期的な見直しの対象となります。AI オフィスは、AI 技術の進歩、社会の変化、新たなシステムリスクを反映するために、本規範の更新を奨励し、促進することができます。

g) 署名者は、本規範が汎用 AI モデルに関する統一された EU 標準が採用されるまでの橋渡しとして機能することを認識しています。将来の標準への段階的な移行を促進するために更新が必要になる場合があります。

h) 署名者は、本規範に特定の測定基準、サブ測定基準、主要業績評価指標（KPI）が存在しないとしても、汎用AIプロバイダーが免除されるわけではないことを認識する。
システムリスクのあるモデルには、潜在的なシステムリスクが発生したときにそれに対処し、軽減する責任があります。

i) AI オフィスと署名者は、AI 環境における新たな課題と機会に対処するために、汎用 AI モデルの提供者、研究者、規制機関間の協力を促進するために協力するものとします。

規範の目的は次のとおりです。

I. 汎用AIモデルの提供者は、義務を効果的に遵守できます。行動規範は、提供者に対してコンプライアンスを証明する方法を明確にする必要があります。また、行動規範は、AIオフィスが、第56条に従って、行動規範に依拠してコンプライアンスを証明することを選択した提供者のコンプライアンスを評価できるようにする必要があります。これには、特に汎用AIモデルの開発と展開の傾向を十分に把握できるようにすることが含まれます。

最も先進的なモデルの。

II. 汎用AIモデルの提供者は、AIバリューチェーンに沿って汎用AIモデルを十分に理解しておくことで、そのようなモデルを下流の製品に統合できるようにし、AI法やその他の規制に基づくその後の義務を果たすことができます（第53条および第101条参照）。

III. 汎用AIモデルの提供者は、著作権および関連する権利に関するEU法を効果的に遵守することができます（第53条および第106条を参照）。

IV. システムリスクを伴う汎用AIモデルの提供者は、システムリスクを伴う汎用AIモデルの開発、市場投入、または使用から生じる可能性のあるシステムリスク（その発生源を含む）を、EUレベルで継続的に効果的に評価し、軽減することができます（第55条および第114条参照）。

II. 汎用AIプロバイダーに対する規則

モデル

一方：

a) 署名国は、汎用AIプロバイダーの特別な役割と責任を認識している。
AIバリューチェーンに沿ったモデルは、提供されるモデルがさまざまな下流システムの基礎となる可能性があるため、下流プロバイダーによって提供されることが多く、下流プロバイダーは、モデルを自社製品に統合できるようにし、AI法に基づく義務を果たすために、モデルとその機能を十分に理解する必要があります。[1]

b) 署名者は、汎用 AI モデル、特にテキスト、画像、その他のコンテンツを生成できる大規模な生成 AI モデルが、ユニークなイノベーションの機会を提供すると同時に、アーティスト、著者、その他のクリエイター、そしてそのクリエイティブ コンテンツの作成、配布、使用、消費の方法に課題をもたらすことを認識しています。また、著作権で保護されたコンテンツの使用には、関連する著作権の例外および制限が適用されない限り、関係する権利者の許可が必要であることも認識しています。2

c) 署名国は、モデルの変更または微調整の場合には、比例性を保つためにプロバイダーの義務はその変更または微調整に限定されるべきであると認識している。3

d) AI法および本規約は、EUおよび国内法で定められた規則に影響を与えるものではなく、本規約は特にEU著作権法に従って解釈されるものとします。指令(EU)2019/790は、一定の条件下でテキストおよびデータマイニングの目的で、作品またはその他の主題の複製および抽出を許可する例外および制限を導入しました。

これらの規則に基づき、権利者は、科学的研究の目的で行われる場合を除き、テキストやデータのマイニングを防ぐために、作品やその他の主題に対する権利を留保することを選択できます。
適切な方法で権利が明示的に留保されている場合、汎用 AI モデルの提供者は、そのような作品に対してテキストおよびデータマイニングを実行する場合、権利保有者から許可を得る必要があります。4

e) 署名国は、AI法第53条(1)(c)項に基づき、汎用AIモデルをEU市場に投入するすべてのプロバイダーは、著作権および関連する権利に関するEU法を遵守するためのポリシーを策定する義務があり、特に、それらの汎用AIモデルのトレーニングの基礎となる著作権関連行為が行われる管轄区域にかかわらず、指令(EU)2019/790第4条(3)項に従って表明された権利の留保を特定し、最先端の技術を含む手段を通じて遵守する義務があることを認識している。5このセクションの第2章は、さまざまな権利と正当な権利の間で公平なバランスを保ちながら、著作権の遵守と透明性を確保するための堅牢なフレームワークを設定することにより、この義務の適切な適用6に貢献することを目指している。

1リサイタル101。
2リサイタル105。
3リサイタル109。
4リサイタル105。
5リサイタル106。
6第56条(1)

問題となっている利益7。これらの措置は、新興企業を含む中小企業の利益も適切に考慮しています。

したがって、この規範の署名者は以下のことを約束します。

7 CFEUおよびCJEUの2008年1月29日の判決、Promusicae（C-275/06、ECR 2008 p. I-271)ECLI:EU:C:2008:54、68項、2014年3月27日の判決、UPC Telekabel Wien（C-314/12)第17条（2）、第16条および第13条を参照
ECLI:EU:C:2014:192、パラ46; 2022 年 4 月 26 日、ポーランド / 議会および評議会の判決 (C-401/19、デジタルレポートに掲載) ECLI:EU:C:2022:297、パラ66。

透明性

法律文書

第53条(1)(a)：「汎用AIモデルの提供者は、モデルのトレーニングおよびテストのプロセス、評価の結果を含む技術文書を作成し、最新の状態に保つものとし、これには、AIオフィスおよび各国の所管当局の要請に応じて提供することを目的として、少なくとも附属書XIに記載されている情報が含まれるものとする。」

第53条(1)(b)：「汎用AIモデルの提供者は、汎用AIモデルを自社のAIシステムに統合しようとするAIシステム提供者に対し、情報および文書を作成し、最新の状態に保ち、提供するものとする。EU法および国内法に従って知的財産権および企業秘密または営業秘密を遵守し保護する必要性を損なわずに、情報および文書は、(i) AIシステム提供者が汎用AIモデルの能力と限界を十分に理解し、本規則に基づく義務を遵守できるようにすること、および(ii) 少なくとも附属書XIIIに定める要素を含むこと。」

対策1. AIオフィスの文書化

署名者は、要請に応じて AI オフィスおよび各国の所管当局に提供するために、以下の表に記載されているモデルの技術文書を作成し、最新の状態に保つことを約束します。署名者は、公の透明性を高めるために、記載されている情報の全部または一部を一般に公開できるかどうかを検討することが推奨されます。

対策2. 下流プロバイダー向けの文書化

署名者は、汎用 AI モデルを自社の AI システムに統合する予定の AI システム プロバイダーに対して、以下の表に記載されている情報と文書を作成し、最新の状態に保ち、提供することを約束します。署名者は、記載されている情報を、完全に公開された状態で透明性をもって開示できるかどうか検討することが推奨されます。

または で に の に 前進 公共

AI法参照	必要な情報の詳細	AIの場合 事務所および 国の管轄当 局	下流 プロバイダー向け
附属書XI §1 1.および附 属書XII 1.	一般情報: 署名者は、汎用AIプロバイダーに関する一般情報を詳細に記載する必要がある。 モデルとモデル自体に関する情報は、モデル名、バイナリが配布されている場合は安全なハッシュ、サービスの場合は TLS/SSL 証明書などによるモデルの出所と信頼性の証拠、モデルの開発者と所有者の正式商号（両者が異なる場合）、モデル ファミリの商号、提出された各モデル バージョンの一意の名前など、モデルを明確に識別して特徴付けるために必要です。		

附属書XI §1 1.(a)および附属書XII 1.(a)	意図されたタスクと、それを統合できるAIシステムの種類と性質: 署名者は、意図されたタスクと制限または禁止されたタスクの説明を提供する必要があります。この説明には、汎用AIが統合されるAIシステムの種類と性質も含める必要があります。 モデルを統合できる対象には、高リスク AI アプリケーション（付録 III に指定）も含まれます。		
附属書XI §1 1.(b) および附属書XII 1.(b)	許容される使用ポリシー: 署名者は、プロバイダー間の一般的な慣行に基づいて、許容される使用ポリシー (AUP) の詳細を提供する必要があります。有効な許容される使用ポリシーには、少なくとも付録で定義されている必須要素が含まれている必要があります。署名者は、最新の許容される使用ポリシーのアクティブな URL を公開する必要があります。		
附属書XI §1 1.(c) および附属書XII 1.(c)	リリース日と配布方法: 署名者は、一般公開の配布方法の最新リストとともにリリース日を提供する必要があります。 目的AIモデル。署名者は、最新の配布方法についてのアクティブなURLを公開するものとします。		
付属書XII 1.(d)	モデルと外部ハードウェアまたはソフトウェアとの相互作用: 署名者は、モデルがハードウェアおよびソフトウェアとどのように相互作用するかについて、どのハードウェアおよびどのソフトウェアがモデルの一部ではないかを指定して、ドキュメントを提供する必要があります。署名者は、必要なソフトウェアおよび/またはハードウェアのバージョン依存関係を公開するものとします。		
附属書XII 1.(e)	該当する場合、関連ソフトウェアのバージョン: 署名者は、汎用 AI モデルを使用するために必要な関連ソフトウェアのバージョンに関する詳細を提供する必要があります。署名者は、必要なソフトウェアのバージョン依存関係を公開するものとします。		
附属書XI §1 1.(d)および附属書XII 1.(f)	アーキテクチャとパラメータの数: 署名者は、モデル アーキテクチャ、モデルの種類、適切な場合はコンテキスト サイズ、モデル パラメータの合計数、および推論中にアクティブなパラメータの数について説明する必要があります。		
	署名者は、モデルのレイヤーの数や種類など、モデル アーキテクチャに関する詳細情報を提供する必要があります。		
附属書XI §1 1.(e) および附属書XII 1.(g) および 2.(b)	入力と出力の様式と形式: 署名者は、該当する場合、入力と出力の様式および関連するコンテキストの制限を詳細に記述する必要があります。		
附属書XI §1 1.(f) および附属書XII 1.(h)	ライセンス: 署名者は、プロバイダー間の一般的な慣行に基づいて、ライセンスのコア要素を詳細に記述する必要があります。これには、公開される資産（データ、モデルの重みなど）に関する情報と、ライセンスの義務が含まれます。		

	使用、変更、配布の条件。署名者は、最新のライセンスの有効な URL を公開するものとします。		
附属書XI §1 2.(a) および附属書XII 2. (a)	AIシステムへの統合のための技術的手段: 署名者は、汎用 AI モデルを AI システムに適切に統合するために必要な技術文書、インフラストラクチャ、およびツールを詳細に規定する必要があります。 署名者は、必要なソフトウェアおよび/またはハードウェアのバージョン依存関係を公開するものとします。		
附属書XI §1 2.(b)	設計仕様とトレーニング プロセス: 署名者は、モデル トレーニングの中核要素 (トレーニング段階、最適化される目的、最適化の方法、制約など)、設計上の決定に関連する根拠と仮定、およびその他のトレーニングの詳細を詳述する必要があります。		
附属書XI §1 2.(c) および附属書XII 2. (c)	トレーニング、テスト、検証に使用されるデータに関する情報: 署名者は、データ取得方法、各データ取得方法の具体的な情報 (Web クロール、データ ライセンス、データ注釈、合成生成データ、ユーザー データなど)、データ処理の詳細 (有害データやプライベート データがフィルタリングされるかどうか、フィルタリングされる場合の方法など)、およびモデルのトレーニング/テスト/検証に使用されるデータに関する具体的な情報 (さまざまなデータ ソースから取得されるデータの割合、トレーニング、テスト、検証データの主な特性など) を詳述する必要があります。		
	署名者は、各データ モダリティ (テキスト、画像、ビデオなど) のトレーニング、テスト、検証データのサイズ (データ ポイントの数) と、データ ソースの不適切さやデータ内の偏りを検出するために使用される方法についてさらに詳しく説明する必要があります。		
附属書XI §1 2.(d)	計算リソース: 署名者は、比較可能で検証可能な文書化を可能にするために測定および計算方法を詳述するために AI 法第 97 条に従って採択された委任行為と一致して、モデルのトレーニングおよび推論に使用される計算リソース (汎用 AI モデルのトレーニングおよび推論に必要なハードウェア ユニットの数と種類、トレーニング プロセスの期間、FLOP 数など) を詳述する必要があります。		
附属書XI §1 2.(e)	エネルギー消費: 署名者は、比較可能で検証可能な文書化を可能にするために測定および計算方法を詳述するために AI 法第 97 条に従って採択された委任行為と一致して、エネルギー消費を評価するために使用する情報と方法 (ハードウェア プロバイダー、ハードウェアに関連する場所とエネルギー源、消費されたエネルギー、発生する推定排出量など) を詳述する必要があります。		

第53条（1）（a）	テストプロセスとその結果: 署名者は、テストが実施されていない場合も含め、汎用 AI モデルのテストプロセスを詳しく説明する必要があります。 これらの詳細には、適切な解釈を確実にするために、実行されたテストの説明とこれらのテストの結果を含める必要があります。		

未解決の質問

上記の表に記載されている項目について、規範ではどのように詳細を規定すべきでしょうか?

付録: 利用規定の必須要素

許容使用ポリシー (AUP) は、サービスまたはテクノロジーの使用方法を概説した一連のルールとして定義されます。これは、許容される動作と許容されない動作に関するガイドラインをユーザーに提供するドキュメントです。AUP は、汎用 AI モデルの使用法と機能を説明する署名者の資料と一致している必要があります。署名者は、汎用 AI モデルに関連する必要なすべての情報を下流プロバイダーと共有し、下流プロバイダーが AI システムの使用目的のタスクまたはユースケースに適用される既存の規制に準拠できるようにする必要があります。

AUP には、少なくとも次の内容が含まれている必要
があります。• AUP が存在する理由を説明する目的の記述。• ポリシー
の適用対象と対象となるリソースを定義する範囲。• 主な使用目的とユーザー。

• 許容される使用方法、高リスクのAIアプリケーションを含む、許可されているアクティビティとタスクを一覧表示する
（附属書IIIに規定されるもの)がある場合、モデルはそれに統合されることが意図されている。
• 許容されない使用法、禁止されている行為の詳細。 • セキュ
リティ対策、一般的なユーザーが遵守すべきセキュリティプロトコルの説明を含む。
AI システムが従わなければならない目的。
•監視とプライバシー、汎用AIプロバイダーがモデルの使用とユーザーのプライバシーへの影響を監視する理由と方法を説明する。
• 規則に従わなかった場合にユーザー権限を停止または
取り消すための警告プロセスと基準。
AUP;
• ユーザーアカウントの終了基準と適用法規制への参照
執行;
• 承認。下流プロバイダーが AUP を読んで理解し、遵守することに同意したことを確認する必要があります。

著作権に関する規則

法律文書

第53条(1)(c)：「汎用AIモデルの提供者は、著作権および関連する権利に関するEU法を遵守し、特に、最先端の技術を通じて、指令(EU)2019/790の第4条(3)に従って表明された権利の留保を特定し、遵守するためのポリシーを導入しなければならない。」

対策3. 著作権ポリシーを導入する

署名者は、著作権および関連する権利に関する EU 法に準拠するためのポリシーを導入することを約束します。

対策3を満たすために：

サブ措置3.1. 著作権ポリシーの策定と実施

署名者は、本章の規定に沿って、著作権および関連する権利に関する欧州連合法に準拠するための内部ポリシーを作成し、実施するものとします。このポリシーは、本規定の対象となるあらゆる汎用 AI モデルのライフサイクル8 全体をカバーするものとします。汎用 AI モデルの変更または微調整の場合、汎用 AI モデルの提供者の義務は、AI オフィスのガイダンスに従って、新しいトレーニング データ ソースを含むその変更または微調整のみに関係します。

[9] 署名者は、組織内で実施の責任を割り当て、
このポリシーの監督。

サブ措置3.2. 上流著作権コンプライアンス

署名者は、汎用 AI モデルの開発のためのデータセットの使用について第三者と契約を締結する前に、合理的な著作権デューデリジェンスを実施します。特に、署名者は、指令 (EU) 2019/790 の第 4 条 (3) に従って表明された権利留保を、最先端の技術を含む手段を通じて第三者がどのように特定し遵守したかについて、第三者に情報を要求することが推奨されます。

サブ措置3.3. 下流著作権コンプライアンス

署名者は、汎用AIモデルが統合されている下流のシステムまたはアプリケーションが著作権を侵害する出力を生成するリスクを軽減するために、合理的な下流の著作権対策を実施する。10下流の著作権ポリシーでは、署名者が独自の汎用AIモデルを自社のAIシステムに垂直統合するかどうか、または契約関係に基づいて汎用AIモデルが別のエンティティに提供されるかどうかを考慮する必要がある。11 特に、署名者は、汎用AIモデルの過剰適合を避け、汎用AIモデルを別のエンティティに契約で提供する結論または有効性を、

8リサイタル65。
9リサイタル109。
10第3条(1)及び第3条(63)。
11序文97および第3条(68)。

当該事業体が、保護対象作品と同一または類似と認識される成果物の繰り返し生成を回避するために適切な措置を講じることを約束した場合。このサブ措置は中小企業には適用されません。

対策4. TDM例外の制限の遵守

署名者は、汎用AIモデルの開発のために指令（EU）2019/790の第2条（2）に従ってテキストおよびデータマイニングを行う場合、著作権で保護されたコンテンツへの合法的なアクセス権を確保し、指令（EU）2019/790の第4条（3）に従って表明された権利留保を特定し、遵守することを約束します。

対策4を満たすために：

サブ対策 4.1. Robots.txt を尊重する

署名者は、ロボット排除プロトコル (robots.txt) に従って表現された指示を読み取り、それに従うクローラーのみを使用します。

サブ対策4.2. 発見可能性に影響なし

2022/2065年EU規則第3条(j)に定義されているオンライン検索エンジンも提供している、またはそのようなプロバイダーを管理している署名者は、ロボット排除プロトコルに従って表明されたクローラー排除が検索エンジンでのコンテンツの検索可能性に悪影響を及ぼさないように適切な措置を講じます。

サブ措置4.3. その他の適切な手段に関する最善の努力

署名者は、オンラインで公開されるコンテンツの場合、指令（EU）2019/790の第4条（3）に従って、ソースレベルおよび/または作品レベルで権利留保を表現するために、広く使用されている業界標準に従って、他の適切な機械可読手段を特定し、遵守するために最大限の努力をします。
特に、署名国は、集約レベルで権利留保の表現を可能にする、広く採用されているツールを実装することが推奨されます。

サブ措置4.4. 権利留保基準の共同開発への取り組み

欧州委員会の招待を受けて、署名者は、影響を受ける権利保有者を十分に代表する団体、および標準化団体などのその他の関連利害関係者と真摯な協議を行い、指令 (EU) 2019/790 の第 4 条 (3) に従って権利留保を表明し、そのような権利留保を特定して遵守するための相互運用可能な機械可読標準を開発します。欧州委員会は会議を招集して議長を務め、必要に応じて関連利害関係者および汎用 AI プロバイダーと協議した後、プロバイダーが尊重することが期待される最先端のソリューションに関する情報を発行する場合があります。このサブ措置は中小企業には適用されません。ただし、中小企業はこれらの協議に自主的に参加することができます。

サブ措置4.5. 著作権侵害ウェブサイトのクローリング禁止

署名者は、欧州委員会の偽造•著作権侵害監視リストに掲載されているウェブサイトを除外するなど、海賊版ソースをクローリング活動から除外するための合理的な措置を講じる。署名者はまた、

設立された管轄区域内の関連公的機関が発行する類似の除外リストに従うことが推奨されます。

対策5. 透明性

署名国は、著作権および関連する権利に関する欧州連合の法律を遵守するために採用する措置について十分な透明性を確保することを約束します。

対策5を満たすために：

サブ措置5.1. 権利留保の遵守に関する公開情報

署名者は、指令（EU）2019/790の第4条（3）に従って表明された権利留保を特定し遵守するために採用する措置に関する適切な情報を、可能な限り多くのEU市民に広く理解される言語で公表するものとする。当該情報は、各署名者のウェブサイトで容易にアクセスでき、最新の状態に保たれるものとする。

サブメジャー 5.2 クローラー名と robots.txt の機能

前述のサブ措置による情報には、少なくとも、署名者が本規約の対象となる汎用 AI モデルの開発に使用するすべてのクローラーの名前と、クロール時を含む関連する robots.txt 機能が含まれます。

サブ措置5.3. 単一の連絡窓口と苦情処理

署名者は、権利者が電子的手段で直接かつ迅速に連絡できるよう、単一の連絡先を指定することが推奨されます。特に、権利者とその代表者（集団管理団体を含む）が、その作品やその他の保護対象が汎用 AI モデルの開発に利用されることに関して苦情を申し立てられるようにし、適切な苦情処理手順を実施することが推奨されます。

サブ措置5.4 データソースと承認の文書化

AI事務局が署名国が著作権および関連する権利に関するEU法を遵守するための政策を策定する義務を果たしているかどうかを監視できるようにするために、[13] 署名者は、
トレーニング、テスト、検証に使用されるデータソースに関する情報と、汎用 AI の開発のために保護されたコンテンツにアクセスして使用するための承認に関する情報を AI オフィスに提供し、要求に応じて提供します。

12第89条(1)
13 108条の理由及び53条(1)項。

III. システムリスクの分類

一方：

a) 署名国は、システミックリスクの分類には種類、性質、発生源が含まれることを認識している。
システムリスクの。
b) 署名国は、この分類法が開発されており、疑義がある場合には、AI法第3条(2)に定義されている各リスクの重大性と確率、およびAI法第3条(65)に定義されているシステミックリスクの定義に照らして解釈されるべきであることを認識している。
c) 署名国は、システミックリスクの分類は網羅的なものではなく、科学の進歩や社会の変化を反映して時間の経過とともに変化する可能性があることを認識している。
d) 署名国は、第3節および第4節が一般的に汎用AIを指していることを認識している。
AIシステムではなくモデルがリスクの一部であるが、リスクは、汎用AIモデルをAIシステムにどのように展開できるかを考慮することで、最も適切に特定、評価、軽減できることが多い。汎用AIモデルプロバイダーがAIシステムも開発•運用している場合、

システムリスクを伴う汎用 AI モデルに基づいて、これらのシステムを考慮してリスク評価と軽減 (安全性とセキュリティのフレームワークで説明) を実行します。

したがって、この規範の署名者は以下のことを約束します。

対策6. 分類

署名者は、システムリスクの評価と軽減の基礎として、このシステムリスクの分類の要素を活用することを約束します。

6.1. システムリスクの種類

署名者は、以下のものをシステムリスクとして扱います。

•サイバー攻撃 :脆弱性の発見や、
搾取。
•化学、生物、放射線、核のリスク:軍民両用科学は、兵器の開発、設計、取得、使用などを通じて、化学、生物、放射線、核兵器による攻撃を可能にするリスクがあります。

•制御不能 :強力な自律型汎用車両を制御できないことに関連する問題
AI モデル。
• AI 研究開発のためのモデルの自動使用:これにより AI 開発のペースが大幅に加速し、システムリスクを伴う予測不可能な汎用 AI モデルの開発につながる可能性があります。

•説得と操作:選挙への干渉、メディアへの信頼の喪失、知識の均質化や単純化など、民主的な価値観や人権に対するリスクを伴う大規模な説得と操作、および大規模な偽情報や誤情報の促進。

•大規模な差別 :個人、コミュニティ、または
社会。

署名者は、例えば、重大な事故、大規模なプライバシー侵害や監視、また、汎用 AI モデルが公衆衛生、安全、民主的プロセス、公共および経済の安全、重要なインフラ、基本的権利、環境資源、経済の安定、人間の行為、あるいは社会全体に大規模な悪影響を及ぼす可能性のあるその他の方法などを考慮して、上記に挙げたもの以外にもさらなる体系的なリスクを特定する可能性があります。

未解決の質問

- リスクが危険であるかどうかを判断する際に考慮すべき考慮事項や基準は何ですか？
  システムリスクですか?
- これらの考慮事項や基準に基づいて、どのリスクを優先して追加すべきか
  システミックリスクの主な分類は何ですか?
- システムリスクの分類では、AI によって生成された児童性的虐待資料や同意のない親密な画像にどのように対処すべきでしょうか?

6.2. システムリスクの性質

システミックリスクの性質とは、リスクを評価し軽減する方法に影響を与えるリスクの主要な属性を指します。署名者は、システミックリスクの性質の特に関連する以下の側面と、網羅的でも相互に排他的でもない各側面の例を考慮します。

- 起源:モデルの機能、モデルの配布
- リスクを引き起こす主体:国家、グループ、個人、自律型AIエージェント、なし（例: 明確な主体がいない）
  識別できる）
- 意図:意図的、非意図的（不一致を含む）
- 新規性:前例のない、前例のない
- 確率と重大度の比率:影響度が低い、確率が高い、影響度が高い、確率が低い、
  予想される影響 • リスクが顕在化する速度:徐々に、突然に、継続的に変化する • リスクが顕在化する際の可視性:明白(オープン)、隠れた (隠れた) • イベントの進行:線形、再帰的 (フィードバックループ)、複合的、カスケード的 (連鎖反応)

6.3. システムリスクの原因

リスクの源泉は、「リスクの要因」または「リスクの推進要因」とも呼ばれ、単独または組み合わせでリスク（例：モデルの盗難や広範囲にわたるサイバー脆弱性）を引き起こす要素（例：イベント、コンポーネント、アクター、およびそれらの意図または活動）です。署名者は、以下を特に関連があると考えています。
システムリスクの原因:

6.3.1. 危険なモデル機能これらは、システム
リスクを引き起こす可能性のあるモデル機能です。署名者は、これらの多くが
有益な使用には機能も重要です。これには次のものが含まれます。

- サイバー攻撃能力、化学•生物•放射線•核兵器（CBRN）能力、
  兵器の取得または拡散能力

• 自律性、拡張性、新しいタスクを学習する適応性 • 自己複製、自己改善、およ
び他のモデルをトレーニングする能力 • 説得、操作、欺瞞 • 長期的な計画、予測、戦略策定 • 状
況認識

6.3.2 危険なモデル傾向これらは、システム
リスクを引き起こす可能性のある、能力を超えたモデル特性です。これには以下が含まれます。

• 人間の意図や価値観との不一致 • 欺く傾向

• バイアス
• 作話
• 信頼性とセキュリティの欠如 • 「目標追求」、
目標変更への抵抗、および「権力追求」 • 他のAIモデル/システムとの「共謀」

6.3.3 モデルのアフォーダンスと社会技術的コンテキストこれらは、モデルの機能や
傾向を超えた要因であり、モデルがもたらす体系的リスクに影響を与える可能性があります。体系的リスクを伴う汎用 AI モデルの特定の入力、構成、およびコンテキスト要素が含まれます。これには次のものが含まれます。 • ガードレールを外す可能性 • ツールへのアクセス (他のモデルを含む) • モダリティ (新しいモダリティと複合モダリティを含む) • リリースおよび配布戦略 • 人間による
監視 • モデルの流出 (例: モデルの漏洩/盗難) • ビ
ジネス ユーザーの数とエンド ユーザーの数

• 攻撃と防御のバランス、悪意のある行為者の数、能力、悪用する意欲など
モデル
• 社会的脆弱性または適応性 • 説明可能性または透
明性の欠如 • 技術の準備状況（つまり、特定のアプリケ
ーションコンテキスト内での技術の成熟度）

データ、モデル、推論の使用におけるフィードバックループ

IV. 汎用プロバイダーの規則

システムリスクを伴うAIモデル

説明ボックス

対策、サブ対策、KPI は比例的である必要があります。特に、特定のプロバイダー、特に AI 開発の最前線にいる企業よりも資金力の少ない中小企業や新興企業の規模と能力、および適切な場合には比例の原則を反映し、利益とリスクの両方を考慮したさまざまな配布戦略 (例: オープンソース) に合わせて調整する必要があります。

現在の草案は、システミックリスクを伴う汎用モデルとその提供者の数は少数であると想定して作成されています。これらの数が増えた場合、例えば、最も大きなシステミックリスクをもたらすモデルに主に焦点を当てることを目的とした、より詳細な階層化対策システムを導入するなど、将来の草案では大幅な変更が必要になる可能性があります。

すぐ下の「一方」の部分は、セクション IV の前文です。ここでは、高レベルの原則が、対策、サブ対策、および KPI の解釈を導きます。

最後に、これはまだ最初の草稿です。皆様のご意見をお待ちしております。
関連する未解決の問題についてはご意見をお待ちしていますが、草案の他の部分についてもご意見をお待ちしています。また、さまざまなビジネス モデルや展開戦略に合わせて、対策をよりバランスよく、より適切にするための提案も歓迎します。

ワーキンググループ2、3、4の議長および副議長。

法律文書

第55条(1)：「第53条および第54条に列挙された義務に加えて、汎用通信事業者は、

システムリスクを伴う AI モデルは、次の要件を満たす必要があります。

(a) システムリスクを特定し、軽減することを目的として、モデルの敵対的テストを実施し、文書化することを含め、最新技術を反映した標準化されたプロトコルとツールに従ってモデル評価を実施する。(b) システムリスクを伴う汎用AIモデルの開発、市場投入、または使用から生じる可能性のあるシステムリスク（その
発生源を含む）を、欧州連合レベルで評価し、軽減する。

（c）重大な事件及びそれに対処するための可能な是正措置に関する関連情報を追跡し、記録し、AIオフィス及び適切な場合には国の所管当局に遅滞なく報告する。

（d）汎用AIモデルに対する適切なレベルのサイバーセキュリティ保護を確保する。
「体系的なリスクとモデルの物理的インフラストラクチャ」

一方：

a) 署名者は、システミックリスクを伴う汎用 AI モデルの提供者は、モデルのライフサイクル全体にわたって適切な措置を講じ、AI バリューチェーンの関連関係者と協力し、機能の向上や新たな可能性を考慮して定期的に実践を更新することで、リスク管理が将来にわたって有効であることを確保しながら、システミックリスクを継続的に評価および軽減する必要があることを認識しています。[14]

b) 署名者は、システミックリスクのある汎用AIモデルが(i) 重大なシステミックリスクをもたらす可能性が高い場合、(ii) 機能と影響が不確実である場合、または(iii) プロバイダーに関連する専門知識が不足している場合、詳細なリスク評価、緩和策、および文書化が特に重要であることを認識しています。逆に、新しい汎用AIモデルが、すでに安全に展開されているシステミックリスクのある汎用AIモデルと同じ大きな影響を与える機能を発揮すると信じる十分な理由があり、重大なシステミックリスクが顕在化せず、適切な緩和策の実施が十分である場合、より包括的な対策の必要性は低くなります。

規模や能力の異なるプロバイダー間で利用可能なリソースを共有し、比例の原則を認識した上で、中小企業や新興企業向けに簡素化されたコンプライアンス方法が適切な場合に提供されます。

c) 署名者は、体系的リスクの評価と軽減を支援する上で、優れた専門知識を持ち、適切な立場にある組織が多岐にわたることを認識しています。

d) 署名者は、多くのリスク評価方法には多大な作業量とコストが伴うことを認識しています。
彼らは、例えば評価やベストプラクティスを共有するなどして、お互いに「負担を分担する」ことを奨励し合っている。
またはインフラストラクチャ、または適切な場合には、業界団体の支援を受けて資格のあるサードパーティプロバイダーと連携することによって実現します。

e) 署名者は、疑義がある場合には、以下の点を考慮して、措置、サブ措置、KPIを解釈するものとする。
システムリスクの効果的な評価と軽減。

したがって、この規範の署名者は以下のことを約束します。

対策7. 安全とセキュリティの枠組み

署名国は、安全とセキュリティの枠組み（SSF）を採用、実施、および利用可能にすることを約束する。この枠組みには、システミックリスクを伴う汎用AIモデルから生じるシステミックリスクを積極的に評価し、比例的に軽減するために遵守するリスク管理ポリシーの詳細が記載されるものとする（第55条（1）参照）。SSFの包括性およびその中のコミットメントは、そのようなモデルの開発から生じると予想されるシステミックリスクの重大性に比例するべきである。当初必要な草案は、

SSF のコンポーネントについては、このセクションの残りの部分で概説します。

14 AI法第114条（「体系的リスクを伴う汎用AIモデルの提供者は、例えば説明責任やガバナンスプロセスなどのリスク管理ポリシーの導入、市販後監視の実施、モデルのライフサイクル全体にわたる適切な措置の実施、AIバリューチェーンに沿った関連関係者との協力などを通じて、体系的リスクを継続的に評価し、軽減する必要がある」）。

汎用AIプロバイダー向けリスク評価

システムリスクを考慮したモデル

対策8. リスクの特定

SSFの一環として、署名者は継続的かつ徹底的にシステムリスクを特定することにコミットしている。
システムリスクを伴う汎用 AI モデルに起因する可能性があります。

措置8を満たすために：

サブ措置8.1. リスクの特定

署名者は、システミックリスクを伴う汎用 AI モデルの提案された開発、市場投入、または使用に特に関連するシステミックリスクを決定し、指定します。この目的のために、署名者は分類法 (セクション III) に記載されているシステミックリスクを使用し、追加のリスクを考慮したり、分類法の他の要素を参照したりする場合があります。

対策9. リスク分析

SSF の一環として、署名者は特定されたシステムリスクへの経路を継続的かつ徹底的に分析することを約束します。

措置9を満たすために：

サブメジャー9.1. 方法論

署名者は、堅牢なリスク分析手法を使用して、システミックリスクを伴う汎用AIモデルの開発と展開がシステミックリスクを生み出す可能性のある経路を特定します。
特定されたリスクと、それらの経路を通じてそのようなリスクが実現する可能性。

サブ指標9.2. システムリスク指標へのマッピング

署名者は、体系的リスクを伴う汎用 AI モデルについて、特定された体系的リスクへの経路を可能にする可能性のある潜在的に危険なモデル機能、傾向、およびその他のリスク源を特定してマッピングし、これらの各要素について体系的リスク指標を提供します。

サブ措置9.3. 深刻度の段階

署名者は、体系的なリスクを伴う汎用AIモデルについて、特定された危険なモデル機能、危険なモデル傾向、およびその他のリスク源を、以下の重大度レベルに分類する。
少なくとも、適切な対策がなければリスクのレベルが許容できないと考えられる深刻度の段階
安全策。

未解決の質問

- 重症度のレベルはどうなるのか？すでに新しい基準や合意はあるか？
  形にする？
- 「重大さ」は「重大さ」のレベルを表現するのに最も良い方法か、それとも「重大さ」と混同してしまう可能性があるか？
  リスクを確率と重大性の組み合わせとして定義しますか?

サブ対策9.4. リスクの予測

署名者は、サブ措置 9.2 に記載されているシステムリスク指標をトリガーするモデルを開発する予定のタイムラインについて、SSF の最善の努力見積もりを含めることになります。

措置10. 証拠収集

SSF の一環として、署名者は、システミック リスクを伴う汎用 AI モデルによってもたらされる特定のシステミック リスクに関する証拠収集の継続的なプロセスに取り組みます。署名者は、予測からクラス最高の評価までさまざまな方法を活用して、モデルの機能、傾向、その他の影響を調査します。

これら

措置10を満たすために：

サブ尺度10.1. モデルに依存しない証拠

署名者は、システミックリスクを伴う汎用AIモデルに該当する場合、文献レビュー、競合他社やオープンソースプロジェクトの分析、一般的な傾向の予測（例：

アルゴリズムの効率、コンピューティングの使用、エネルギーの使用など）、および市民社会、学術界、その他の関連する利害関係者を巻き込んだ参加型の方法。また、スケーリングモデルから能力の向上を予測するスケーリング法則に取り組むこともあります。このセクションのすべての証拠収集と同様に、これは以下の組織と連携して行われる場合があります。
または、資格のある第三者に外注します。

サブ指標10.2. クラス最高の評価

署名者は、体系的リスクを伴う汎用AIモデルの能力と限界を適切に評価するために、クラス最高の評価が実施されることを保証する。これは、体系的リスクを伴うAIモデルのライフサイクルの最も適切な時期に、さまざまな適切な方法論（例えば、Q&Aセット、ベンチマーク、レッドチームおよびその他の敵対的テストの方法、人間の向上研究、モデル生物、シミュレーション、機密資料の代理評価など）を使用して行われ、評価者によって行われる。

関連するリスクに対して適格な（内部または外部の）評価が必要です。これらの評価の深さは、評価対象のリスクと、そのようなモデルがどの程度のリスクを追加するかに関する不確実性に比例します（たとえば、非常に類似したモデルの動作に関する既存の知識により、必要な評価の深さが軽減される場合があります）。

未解決の質問

特定の評価方法が特定のモデルとリスクに適合しているかどうか、また評価が十分に徹底的であったかどうかを決定する要因は何でしょうか。

サブ尺度10.3. 科学的厳密性およびその他の品質要因

署名者は、高い科学的厳密性をもって評価を実施することを確保する。特に重大度の高い体系的リスクについては、資格のある第三者による主要な結果の検証を通じて、さらなる厳密性を達成するものとする（措置 17 を参照）。署名者は、体系的リスクを伴う汎用 AI モデルを適切に評価するために十分な時間、モデル アクセス、計算予算など、厳格な科学的基準に従って作業するために必要なサポートを社内または社外の評価者に提供するとともに、必要に応じて知的財産権と機密ビジネス情報を保護します。

未解決の質問

高度な科学的厳密さはどのように運用されるべきでしょうか? ゴールドスタンダードとは何でしょうか? また、署名国はいつそれを逸脱すべきでしょうか (たとえば、初期の探索的研究を実施する場合)?

サブ尺度10.4. 能力の引き出し

署名者は、モデルの機能を十分に引き出し、機能を過小評価するリスクを最小限に抑えるために、クラス最高レベルの機能引き出し（例 :微調整、迅速なエンジニアリング、スキャフォールディング、コンピューティングおよびエンジニアリング予算）を使用して評価が実行されていることを保証します。

サブメジャー10.5. システムの一部としてのモデル

署名者は、評価によって汎用AIの能力と限界を評価できることを保証する。
システムリスクを伴うモデルは、モデルが使用されることが意図されており、合理的に予見できる将来の AI システムを代表する AI システムだけでなく、モデルがシステムリスクをもたらす可能性が最大限に明らかになる AI システムでも存在します。

未解決の質問

システムリスクを伴う汎用 AI モデルをオープンソース モデルとして、または B2B 顧客に提供する署名者にとって、このサブ措置はどのように促進されるでしょうか?

サブ尺度10.6. 多様な評価と一般化

署名者は、評価がモデルの計画された使用コンテキストとその多様性に一致するようにし、該当する場合は一般化を示すようにします。たとえば、多言語モデルの言語ベースの評価では、英語だけでなく、ヨーロッパの多様性を考慮した多言語評価に重点を置くことができます。

サブ措置10.7. 探索的作業

署名者は、資格のある第三者（以下を含む)によるオープンエンドのレッドチーム演習など、システミックリスクを伴う汎用モデルに関する相当量の調査作業を確実に実施するものとする。
市民社会と学術界の代表者)が参加する。つまり、彼らは
すでに特定されているリスクや機能に関する証拠収集だけでなく、これらの方法を通じて新たなリスクや新たな機能を特定するよう努めます。

サブメジャー10.8. ツールとベストプラクティスの共有

署名者は、クラス最高の安全性評価、ツール、および付随するベストプラクティスを、AI エコシステムの関連関係者が広く利用できるように努めます。特に特定されたケースでは、署名者は、商業的に機密性の高い情報、公共の安全、拡散リスク、および将来の評価の有効性を保護するために、情報の共有を制限する場合があります。

未解決の質問

- 情報共有を促進するためのチャネル、組織、方法はありますか？
  AI 安全性の最先端で現在取り組んでいる研究チームに過度の負担をかけずに、評価、ツール、ベストプラクティスを提供するにはどうすればよいでしょうか。
- この措置は、それほど多くの資金を持っていないスタートアップや中小企業にとって特に有益でしょうか？
  これらのツールとプラクティスをゼロから開発する能力はありますが、それらを使用できるでしょうか?

サブ措置10.9. 結果の共有

署名者は、評価結果を AI オフィスまたは一般の人々と共有する場合、透明性があり比較しやすい形式で共有するものとします。また、実証結果の不確実性や使用した方法の限界についても透明性をもって報告するものとします。

対策11. リスク評価ライフサイクル

署名者は、少なくとも本措置のサブ措置に概説されている段階、および緩和策を実施する前後（措置12に概説されている緩和策の有効性の評価を含む）において、体系的リスクを伴う汎用AIモデルの開発および展開のライフサイクル全体を通じて、継続的にリスクを評価し、証拠を収集することを約束します。

措置11を満たすために：

サブ対策11.1. トレーニング前

署名者は、体系的なリスクを伴う汎用 AI モデルのトレーニング実行を開始する前に、必要に応じて SSF を更新し、署名者の SSF コミットメントに沿って、評価者 (社内および社外) が証拠収集の準備ができていることを確認します。

サブ対策11.2. トレーニング中

署名者は、定期的なマイルストーン（例えば、実効コンピューティングが4倍になるたびに）で証拠を収集し、進行中の安全性とセキュリティレポート（SSR、対策13を参照）をリスクに応じて更新します。ここでのトレーニングは、「大規模なデータコーパスでの事前トレーニング」のみを意味するのではなく、たとえば、監督下での微調整、強化学習フェーズ、または同様のものも含まれます。

モデルを改良する方法。

サブ措置11.3. 展開中

システムリスクを伴う汎用AIモデルの導入中、署名者は
モデルのSSRを再検討し、特に関連する評価を再実行することでリスク評価を再検討します（および/または

少なくとも 6 か月ごとに、または (内部または外部の) 状況に大きな変化が認められた場合や、その他の理由で以前のリスク評価結果に疑問を抱く理由がある場合は、展開中のモデルの監視から得られた証拠も考慮に入れて、常に新しい評価や改善された評価を実施する必要があります。

サブ措置11.4. 導入後のモニタリング

署名者は、導入後のシステムリスクの監視を行う。署名者は、関連する導入後の情報を継続的に収集し、リスク評価に含めるためのメカニズムを確立する。これらのメカニズムは、さまざまなモデルの統合と使用方法によって異なる場合があります（たとえば、有害な出力やアクションのモデルの監視、システムへの影響の調査など）。署名者は、導入後の監視を、モデルを使用する流通戦略や顧客や業界のタイプに合わせて調整します（たとえば、オープンウェイトモデルの場合、ライセンスの遵守を評価したり、

現実世界でのモデルの使用、またはモデルの科学的分析の研究）。モデルプロバイダー自身が
AI システムを導入する場合、これらのモデルをシステムの一部として監視します。

未解決の質問

システムリスクを伴うオープンウェイトの汎用 AI モデルのプロバイダーが、モデルの下流のユーザーに大きな副作用を与えることなく、リリースしたモデルを監視できるようにする方法はありますか (または存在する可能性がありますか)?

一般プロバイダー向けの技術的リスク軽減

システムリスクを考慮したAIモデルの目的

対策12. 緩和策

署名者は、利用可能な場合は AI オフィスのガイダンスに基づいて、各システムリスク指標または重大度の階層から比例して必要な安全性とセキュリティの緩和策までのマッピングを SSF に詳細に記載することを約束します。
マッピングは、少なくとも、システムリスクを許容できないレベル以下に抑えるように設計する必要があり、また、それを超えるリスクを最小限に抑える方法についても説明する必要があります。

措置12を満たすために：

サブ措置12.1. 安全対策

署名者は、システミックリスクを伴う汎用 AI モデルの使用から生じるシステミックリスクを軽減するために実施する安全緩和策を SSF に詳細に記載します。これらの安全緩和策は、システミックリスク指標または重大度の段階に比例する必要があり、(a) モデルの動作変更、(b) システムに展開するためのモデルの周囲に配置される安全対策、(c) システミックリスクを軽減するために他のアクターが利用できる対策またはその他の安全ツールなどが含まれる可能性があります。

サブ措置12.2. セキュリティ緩和策

署名者は、(a) システムリスクのある汎用AIモデルの未公開の重み、および(b) 未公開の関連資産と、そのような未公開モデルのトレーニングまたは使用に必要な情報の保有から生じるシステムリスクを軽減するために実施するセキュリティ緩和策をSSFに詳述する。未公開モデルの場合、これらのセキュリティ緩和策は、展開決定を正当化するのに十分なリスク評価が行われる前の開発段階で適用する必要がある。リリースされたクローズドモデルの場合、これらのセキュリティ緩和策はモデルの展開中および展開後にも適用される必要があるが、公開された重みまたは関連資産を持つモデルには適用する必要はない。さらに、これらのセキュリティ緩和策は、システムリスク指標または重大度の段階に比例する必要があり、(a) 保護を伴う可能性がある。

保管中、移動中、使用中の重量および資産のセキュリティ（必要に応じてハードウェア レベルを含む）(b)重量および資産へのアクセス制御、監視、および強化されたインターフェイス、(c)継続的なセキュリティ レッドチーム演習および認定セキュリティ レビューによる保証、および (d)内部脅威のスクリーニング。

未解決の質問

- サイバーセキュリティと情報セキュリティのどのような基準を一般に適用すべきか
  システムリスク指標と重大度の段階に応じて、システムリスクを伴う AI モデルを目的に応じてどのように設計しますか?

- 体系的な汎用AIモデルに対するサイバーセキュリティ基準は、どのように策定されるべきか
  他の分野の既存のサイバーセキュリティ標準とは異なるリスクがありますか?

サブ措置12.3. 制限

署名者は、SSF において、既存の安全性およびセキュリティ緩和策の限界を詳細に記述し、特定のシステムリスク指標または重大度の段階に対してシステムリスクを管理するための適切な緩和策がまだ存在しない場合について述べます。

サブ措置12.4. マッピングの適切性を評価するプロセス

署名者は、SSFにおいて、サブ措置12.1におけるシステムリスク指標または深刻度レベルから安全およびセキュリティ緩和策へのマッピングの継続的な適切性を評価するプロセスを詳述するものとする。
12.2. これは、能力の引き出しやサイバーセキュリティの状況の進歩など、モデルの影響に関連する内部および外部要因の変化に対応するために行うべきであり、措置17で概説されているSSF全体の適切性を評価するための全体的なプロセスを超えて行うべきである。

対策13. 安全とセキュリティに関する報告書

リスク軽減と評価の一環として、措置8.～12.の比較可能で検証可能な文書化を確保するために、署名者は、開発するシステムリスクのある汎用AIモデルについて、安全性とセキュリティに関する報告書（SSR)を作成することを約束する。この報告書は、(a)モデル開発と展開のライフサイクルにおける適切な意思決定ポイントで作成され、(b)モデルのリスクと軽減評価の詳細が記述されるものとする。
モデルであり、(c)モデルの開発および展開の決定の基礎となる。

措置13を満たすために：

サブ措置13.1. 比例性

署名者は、SSR の (a) 包括性と詳細レベル、(b) 開発および展開ライフサイクルにおけるタイミング、(c) 外部からの入力と精査のレベルがすべて、評価対象のモデルに関連するシステムリスク指標または深刻度の段階に比例していることを保証します。

サブ措置13.2. リスク評価の結果

署名者は、措置 8 ～ 11 に従って、緩和策が実施される前と実施された後の両方で、モデルに対して実施されたリスク評価の結果を SSR に詳述します。

サブ措置13.3. 安全緩和評価の結果

署名者は、サブ措置 12.1 に従って実施された安全緩和策の有効性の評価結果を SSR で詳細に記載します。

サブ措置13.4. セキュリティ緩和評価の結果

署名者は、サブ措置 12.2 に従って実施されたセキュリティ緩和策の有効性の評価結果を SSR で詳細に記載します。

サブ指標13.5. 費用便益分析

署名者は、配備手続きを正当化するために使用された費用便益分析をSSRに詳細に記載する。
サブ措置14.2に従って。

サブ措置13.6. 方法論に関する十分な詳細

署名者は、SSRが独立した評価を可能にするのに十分な科学的詳細を含むことを保証する。
サブ措置13.2～13.5の結果、証拠、分析を生成するために使用された方法（サブ措置10.3も参照）。

サブ措置13.7. レビュー

署名者は、内部（またはより重大度の高い外部）レビューの結果をSSRに詳細に記載する。
サブ措置13.2～13.6で提供された結果

サブメジャー13.8. 同等性

署名者は、AI オフィスと共有される SSR が、開発または展開の決定のために社内で使用されるものと同じであることを保証します。

対策14. 開発と展開の決定

不十分な安全性とセキュリティ対策によるリスクを軽減するために、署名者は汎用AIモデルの開発と展開を進めるか否かを決定するプロセスを確立することを約束する。
システムリスクを伴う。このプロセスは署名者のSSFに記載され、SSRで提示された結果と分析に基づくものとする。

措置14を満たすために：

サブ措置14.1. 続行しない条件

署名者は、システミックリスクを伴う汎用AIモデルのさらなる開発と展開が進められない条件、または既存の汎用AIモデルが
安全性とセキュリティの緩和策が実装された後、モデルの SSR に基づいて、体系的なリスクのあるモデルは展開から削除されるか、削除されます。

サブ措置14.2. 手続きの条件

署名者はSSFに開発または展開を継続できる条件を詳細に記載する。
より優れた安全およびセキュリティ緩和策の実施や、システムリスク指標または深刻度の段階に応じた厳密な評価プロセスを伴う費用便益分析の提示など、サブ措置 14.1 に従って進めていない場合。

サブ措置14.3. 外部からのインプットと意思決定

署名者は、開発および展開の決定において、AI オフィスなどの関連する政府関係者を含む外部関係者からの意見や承認が必要となる場合を SSF に詳細に記載します。

プロバイダー向けガバナンスリスク軽減

システムリスクを伴う汎用AIモデル

対策15. システムリスクの所有権

署名者は、システミックリスクを評価し、それに比例して軽減するために、経営幹部や取締役会を含むすべての組織レベルでシステミックリスクに関する適切な所有権を確保することにコミットする（第55条(1)項および第114条参照）。

措置15を満たすために：

サブ措置15.1. 執行レベル

署名者は、システミックリスクに対処するために、経営レベルで責任とリソースを割り当てる。
システムリスクを伴う汎用 AI モデルによって生成されます。

サブ措置15.2. 取締役会レベル

署名者は、リスク委員会を設立するなどして、取締役会レベル（または同等のレベル）で、汎用 AI モデルによって生成されるシステムリスクを監視する責任とリソースを割り当てます。

未解決の質問

- 上記のサブ対策は、プロバイダーの規模やその他の関連する要因に応じて実施されるべきか
  - 措置 15 の遵守として適格とな

る可能性のある例をもっと多く、または他に挙げるべきでしょうか?

対策16. 遵守と適切性の評価

署名国はSSFの遵守と適切性を評価することを約束する（第55条（1)および第114条参照）。

措置16を満たすために：

サブ措置16.1. 定期的なSSF評価

署名者は、毎年、その適切性と遵守状況の評価を実施し、文書化する。
SSF は、計画された活動を検討し、それを理事会または同等の機関に提示します。

未解決の質問

- このような評価で答えるべき具体的な質問はありますか? • この文脈では適切

性はどのように定義されるべきですか?

対策17. 独立した専門家による体系的リスクと緩和の評価

署名者は、特に高重大度層については、必要に応じて、ライフサイクル全体を通じて、体系的リスクを伴う汎用 AI モデルのリスクと緩和策について、独立した専門家による有意義な評価を可能にすることを約束します。このような独立した専門家によるリスクと緩和策の評価には、モデル機能の独立したテスト、収集された証拠のレビュー、体系的リスク、緩和策の妥当性などが含まれます。また、SSF と SSR の独立した専門家によるレビューも含まれる場合があります (第 55 条 (1) および第 114 条の序文を参照)。

未解決の質問

- どのような状況であれば、展開前に、体系的なリスクを伴う汎用 AI モデルについて、独立した専門家による体系的なリスク評価を行うことが適切でしょうか。緩和策の評価についてはどうでしょうか。どのような状況では、逆効果または不必要だと思われますか。 • トレーニングの前またはトレーニング中に、ライフサイクル全体にわたって、独立した専門家をリスク評価に反復的に含めることが適切または推奨される状況はありますか。

- 独立したシステミックリスク評価は、情報セキュリティ、システミックリスクのコンポーネントとドキュメントを含む汎用 AI モデルへのアクセスの深さ、テストの範囲、テスト時間、専門知識、透明性など、関連するシステミックリスクの規模と性質にどのように適応できますか。

- 重大度レベルに応じてどのように対策を講じるべきか？

措置17を満たすために：

サブ措置17.1. 展開前

署名者は、汎用AIを導入する前に十分な独立した専門家によるテストを確実に実施する。
リスクと緩和策をより正確に評価し、外部関係者に保証を提供するために、AI オフィスのガイダンスに従って、AI オフィスや適切な第三者評価者などによって、体系的なリスクを伴うモデルを評価する。これには、署名者が収集した証拠の適切な要素のレビューも含まれる場合があります。

未解決の質問

適切な第三者評価者とはどのような人ですか? 現在の業界の未成熟さを考慮して、どのように規範を起草できますか? AI オフィスがプロバイダー、特に中小企業にリスクと緩和策の独立した専門家による評価を確実に提供できるように支援する方法はありますか?

サブ措置17.2. 展開後

署名者は、体系的なリスクを伴う汎用AIモデルの有意義な独立したテストを可能にする。
導入後、必要に応じて、例えば、モデルのライフサイクル全体にわたるリスクを評価し、導入後の適切な変更を特定する。これには、独立した研究者やAIオフィスを含むその他の関連当事者が、十分なアクセス、リソース、報復のない保証を提供することにより、モデルのリスク、限界、特性を有意義に研究できるようにすることが含まれる可能性がある。
正当な研究活動に反する。

未解決の質問

研究セーフハーバーや脆弱性報告など、独立したテストを容易にするさまざまな手段が適切なのはどのような場合ですか?

措置18. 重大インシデントの報告

法律文書

第55条(1)(c)：「第53条および第54条に列挙された義務に加えて、体系的なリスクを伴う汎用AIモデルの提供者は、重大なインシデントに関する関連情報およびそれに対処するための可能な是正措置を追跡、文書化し、AIオフィスおよび必要に応じて国の所管当局に遅滞なく報告しなければならない。」

署名者は、重大なインシデントがシステムリスクを伴う汎用 AI モデルから発生するものである限り、重大なインシデントを特定して追跡し、関連する情報および可能な是正措置を文書化し、AI オフィスおよび必要に応じて国の所管当局に遅滞なく報告することを約束します。

未解決の質問

- 重大なインシデントとはどのような意味か ?AI法第3条(49)でAIシステムに使用されている定義を本規範で使用すべきか、それともシステミックリスクのある汎用AIモデルには別の定義の方が適切か？ • システミックリスクのある汎用AIモデルはどのような条件下で重大なインシデントであると判断されるべきか？

  間接的に重大な事故の発生につながったことがありますか?
- 自動化または自動化を可能にする適切な技術標準またはベストプラクティスはありますか？

  重大なインシデントの AI オフィスへの報告を効率化しますか?

措置18を満たすために：

サブ措置18.1. 重大インシデントの報告プロセス

署名者は、汎用AIに起因する重大なインシデントやニアミスを特定し、記録し、AIオフィスに報告するためのプロセス（スタッフの指名を含む）を確立する。

システムリスクを考慮したモデル。

サブ措置18.2. 対応準備

署名者は、重大なインシデントに対応するためのプロセスを確立し、重大なインシデントに対応して講じられる可能性のある是正措置を事前に定義し、いつ是正措置を講じることができるかについての説明を含めます。

未解決の質問

- 重大なインシデントが発生した場合、どのような是正措置を講じることができますか? いつ是正措置が適切であるかを規範で規定する必要がありますか?
- オープンウェイトまたはオープンソースの重大なインシデント対応プロセスはどのようなものが適切かプロバイダーですか?

措置19. 内部告発の保護

法律文書

第87条：「指令（EU)2019/1937は、本規則の違反の報告およびそのような違反を報告する者の保護に適用される。」

署名者は、内部告発チャネルを実施し、対象となる個人および活動に対して適切な内部告発保護を提供することを約束します。

措置19を満たすために：

サブ措置19.1. 通知

署名者は、内部告発の苦情を提出できる Al Office メールボックスが稼働していることを条件として、従業員にその旨を積極的に通知します。

未解決の質問

- EU指令2019/1937（「内部告発指令」)には、コードで強調することが重要なのでしょうか?
- 内部告発指令には、規範で明確にしたり、さらに規定したりする必要がある部分はありますか? システムリスクの評価と軽減を可能にするために適切と思われる追加の内部告発措置はありますか?

措置20. 通知

署名者は、汎用 AI モデルがシステミックリスクのある汎用 AI モデルとして分類される基準を満たすモデル、SSF、SSR、および適切な場合には重大なシステミックリスクに関する関連情報を AI オフィスに通知することを約束します。このような通知は、第 78 条に従って提供された情報の機密性を保護する AI オフィスの義務を理解した上で行われます。

措置20を満たすために：

サブ措置20.1. システムリスク通知機能を備えた汎用AIモデル

法律文書

第52条(1)：「汎用AIモデルが第51条(1)(a)に規定する条件を満たす場合、関係プロバイダーは遅滞なく、いかなる場合でも2週間以内に委員会に通知しなければならない。」

「当該通知は、当該要件が満たされた後、または満たされることが判明した後に行われる。当該通知には、当該要件が満たされたことを証明するために必要な情報を含めるものとする。委員会は、通知を受けていないシステムリスクを呈する汎用AIモデルを認識した場合、それをシステムリスクを伴うモデルとして指定することを決定することができる。」

署名者は、トレーニング実行を開始する前に、使用する予定の計算能力の量を見積もり、汎用 AI モデルがシステムリスクのある汎用 AI として分類されるかどうかを AI オフィスに通知します。

未解決の質問

AI オフィスには、汎用モデルが高影響力の機能を備えていると推定されるかどうか (したがって、システムリスクのある汎用 AI モデルとして分類されるかどうか) を決定するための分類基準を更新する権限があります。プロバイダーが新しい分類基準を満たすモデルを AI オフィスにいつ通知すべきかが明確になるように記述するにはどうすればよいでしょうか。

サブ措置20.2. SSF通知

署名者は、AI オフィスが安全性とセキュリティのフレームワークの最新バージョンにアクセスできるようにします。

未解決の質問

このアクセスをどのように促進できるでしょうか?

サブ措置20.3. SSR通知

署名者は、特にシステミックリスクを伴う新しい汎用 AI モデルを市場に投入する前に、関連する決定に先立って AI オフィス SSR を送信します。

サブ措置20.4. 重大なシステムリスクの通知

署名者は、重大なシステムリスクが顕在化する可能性があると信じる強い理由がある場合、AI オフィスに通知します。

未解決の質問

システミックリスクが現実化する可能性があると信じる強い理由は何ですか?

対策21. 文書化

署名者は、システムリスクを伴う汎用 AI モデルのライフサイクル全体を通じて、本規範および AI 法のシステムリスクを伴う汎用 AI モデルに関する規定の遵守に関連する証拠を文書化し、要求に応じてこの情報を AI オフィスと共有することを約束します。

これには、付録XIIIの情報など、体系的リスクを伴う汎用AIモデルの分類に関連する証拠が含まれます。また、AIの基準に準拠していることを証明する文書も含まれます。

附属書XI第2項に概説されている情報に加えて、法律および規範で規定されているSSF、SSR、およびリスク評価中に収集された追加の証拠など（第53条(1)(a)を参照）。

未解決の質問

特に小規模プロバイダーのコンプライアンス コストを削減するために、このような文書の標準化されたテンプレートはどのようなものになるでしょうか。注: 今後の草案では、この措置に基づく文書が合理化され、付録 XI、セクション 1、および付録 XII に詳述されているような他の文書要件と統合されるよう努めます。

措置22. 公的透明性

署名者は、SSF および SSR を公開することにより、下流プロバイダー、AI オフィス、一般市民を含むより広範なエコシステムが、特に AI のリスク評価と軽減の科学がまだ初期段階にあることを考慮して、体系的リスクを伴う汎用 AI から生じる体系的リスクをよりよく理解し、軽減できるようにすることを目的として、適切な公的透明性を提供することを約束します。情報を含めることで体系的リスクが大幅に増大したり、社会的利益に不釣り合いな程度に機密の商業情報が漏洩したりする可能性がある場合は、情報が編集されることがあります。

未解決の質問

- より広範なエコシステムに権限を与えてリスクを評価し、軽減することで、システムリスクが減少するのではなく、どのような種類とレベルの公的透明性が増加するか? • モデルカードとシステムカードの公開が一般的に行われていることを考慮すると、この種の公的透明性はどの程度の負担になるか? こうした負担を軽減する対策を設計できるか?

結論

この規範の最初の草案は、4 つの専門ワーキング グループによる既存のベスト プラクティスの予備的レビュー、約 430 件の提出物からの利害関係者の協議情報、プロバイダー ワークショップからの回答、国際的なアプローチ (G7 行動規範、フロンティア AI 安全コミットメント、ブレッチリー宣言、および関連する政府および標準設定機関からの成果を含む)、そして最も重要な AI 法そのものの結果です。この初期段階では、草案は必然的に高レベルであり、主に規範の基礎となる原則と、提案されている対策およびサブ対策を規定しています。

これはあくまでも最初の草案であり、草案の提案は暫定的なものであり、変更される可能性があることを強調しておきます。したがって、私たちは、この規範の内容をさらに発展させ、更新し、2025 年 5 月 1 日に向けたより詳細な最終版に向けて取り組むにあたり、皆様の建設的な意見をお待ちしています。特に以下の点にご注意ください。

1. この最初の草案は、草案作成計画に定められた6 つの主要な考慮事項に基づいています。i) 欧州連合の原則と価値観との整合性、ii) AI 法および国際的なアプローチとの整合性、iii) リスクへの比例性、iv) プロバイダーの規模と能力への比例性、v) AI 安全エコシステムのサポートと成長、vi) 将来への備えです。これらの原則は、AI 法の目的を推進することを目的としています。

2. 私たちは、市民社会、学界、AI 安全機関、業界など、さまざまな視点を持つ関係者からの意見に基づいて、措置、サブ措置、KPI を包括的に検討、開発、改良する必要があることを認識しています。今後の反復は、前述の起草計画と原則に基づいて行われ、AI 法の条項や説明へのより具体的な言及が含まれる可能性があります。上記の例は予備的な性質のものですが、今後の規範の反復に含めるべき措置、サブ措置、KPI に関する詳細なコメントを歓迎します。また、措置、サブ措置、KPI をさまざまなビジネス モデルや展開戦略に比例してより適切にする方法や、草案で概説した未解決の質問を解決する方法に関する提案も歓迎します。

3. 現在の草案は、システミックリスクを伴う汎用 AI モデルとその提供者の数は少ないという前提で作成されていることに留意します。この前提が誤っていることが判明した場合、将来の草案では、例えば、最も大きなシステミックリスクをもたらすモデルに主に焦点を当てることを目的とした、より詳細な段階的対策システムを導入するなど、大幅な変更が必要になる可能性があります。

この草案に関するご意見は、専用プラットフォーム (Futurium) のフォームから、11 月 28 日木曜日 12:00 CET までに専用のフィードバック ポータルを通じてお寄せください。今後の規範の改訂にご協力できることを楽しみにしています。