

汎用 AI 実務規範

第 1 ドラフト

議長および副議長による序文

我々は、4つの作業部会の議長および副議長として、AI法に基づく汎用AI実務規範（以下、「規範」）の第一次ドラフトをここに提示する。実践規範プレナリーの参加者およびオブザーバーは、11月28日（木）12:00（中央ヨーロッパ時間）までに、専用プラットフォーム（Futurium）上のフォームを通じて、書面によるフィードバックを歓迎する。

本規範の第1ドラフトは、4つの作業部会が緊密に協力することにより、汎用AIモデルのプロバイダと、システムック・リスクを伴う汎用AIモデルのプロバイダに対する主要な検討事項を取り上げている：

- 第1作業部会：透明性と著作権関連規則
- 第2作業部会：システムックリスクのリスク識別とアセスメント
- 第3作業部会：システムックリスクの技術的リスク緩和
- 第4作業部会：システムック・リスクのためのガバナンス・リスク緩和

この最初のドラフトは、さらなる改良のための土台として提示するものである。**作業部会内での議論と利害関係者からの追加的な外部インプットの反復プロセスを経て、対策が追加、削除、修正される可能性がある。**本規範は、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。

当規範は、当規範の指針となる原則と目的を概説した、ハイレベルなドラフト計画を盛り込んだ。最初のドラフトは詳細が軽微であるが、このアプローチは、具体的な下位施策と主要業績評価指標（KPI）に関する徹底的な審議を続ける一方で、最終的な規程の形式と内容について、利害関係者に明確な方向性を示すことを目的としている。また、私たちの審議をより深く理解するために、今後のドラフトで進展が期待される分野のいくつかを強調するための公開質問を加えた。これは、様々な利害関係者が効果的に参加し続けられるよう、フィードバックや提出を導く目的もある。

EUのAI法は2024年8月1日に施行され、2025年5月1日までに規範の最終版を完成させる必要があるとしている。今回発表された最初のドラフトは、2025年以降に開発・発表される次世代モデルにも対応できる「将来を見据えた」規範を提供することを目的としている。2024年10月に始まった我々の作業は、多様な利害関係者からのインプットを統合し、反復的な議論を行ってきた。

本第1ドラフトの策定にあたり、委員長および副委員長は、本規程の適用範囲に含まれる事項に関して、主としてAI法の規定を参考にした。従って、本規程に含まれる文脈や定義に別段の定めがない限り、本規程で使用される用語は、AI法の同一の用語を指す。これには、AI法第56条第1項において、既存の様々な国際的アプローチを考慮するよう指示されていることも含まれる。この最初のドラフトでは、AI法の規定を網羅的に参照していないが、今後の改訂ではそうする予定である。

本規範の第1ドラフトは、産業界、学界、市民社会から数百人の参加者を得た共同作業の成果である。また、AIガバナンス、国際的なアプローチ、連邦法の実践規範（偽情報に関する実践規範など）、作業部会メンバーの専門知識や経験など、発展途上の文献からも情報を得ている。

開発プロセスの主な特徴は以下の通りである：

- マルチステークホルダーによる協議には、これまでに430件近くの応募があった。

- 専門知識、経験、独立性、地理的・性的な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 つの専門作業部会がある。
- 2024 年 10 月から 2025 年 4 月にかけて議論とドラフト会議が開催される。

現在のドラフトを改良し、改善するためには、対外的にも対内的にも、協議と検討のための追加時間が必要となる。独立した委員長および副委員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性のあるものとし、利害関係者がアクセスしやすいものとするよう努力する。皆様の継続的なご協力と建設的なご批判を期待している。

11 月 28 日（木）12:00（中央ヨーロッパ時間）までに、専用プラットフォーム（Futurium）上のフォームを通じて、実務規範プレナリー参加者およびオブザーバーによる書面によるフィードバックを歓迎する。

応援ありがとう！

ヌリア・オリバー	アレクサンダー・パウカート	マティアス・サムワルド	ヨシュア・ベンジオ	マリエ・シャーケ
作業部会 共同議長	第 1 作業部会 議長	第 2 作業部会	第 3 作業部会	第 4 作業部会
リシ・ボマサニ	セリーヌ・カステツ＝ルナル	マルタ・ツイオージ	ダニエル・プリヴィテラ	アンカ・ロイエル
第 1 作業部会 副議長	第 1 作業部会	第 2 作業部会 副議長	第 3 作業部会	第 4 作業部会 副議長
	アレクサンダー・ザッヘル	ニタルシャン・ラジクマール	マルクス・アンデルユング	
	第 2 作業部会 副議長	第 3 作業部会 副議長	第 4 作業部会 副議長	

ドラフト計画と原則

現段階では、この第 1 ドラフトには、最終的に採択される規程の細目は含まれていない。なぜなら、i) 規程の構成と原則に関する幅広い合意を目指していること、ii) この第 1 ドラフトでは、そのような提案に必要なレベルの検討を行って細目の提案を作成するのに十分な時間がなかったこと、iii) 規程の（ドラフト）細目は、継続的に最新の動向を反映させるため、最後に更新される予定だからである。私たちの方向性を示すため、この第 1 ドラフトが全般的にハイレベルであるにもかかわらず、いくつかのコミットメントにおいてより具体的な規定を垣間見ること、将来のドラフトで同様の規定がどのような形式と詳細になるかを示している。本規定におけるコミットメントの構成は、「施策」、「下位施策」、「KPI」の下位階層に従っている。これらのどれかが欠けている場合、特に KPI が欠けている場合、これは最終的な決定ではなく、時間的な制約とこの最初のドラフトのハイレベルな性質によるものである。さらに、この第 1 ドラフトには、規程をどのように見直し、更新していくかという項目がまだない。

また、ドラフトを作成する際に従うべき大まかな原則を以下に示す：

- I. **EU の原則と価値観との整合性** - 施策、下位施策、KPI は、EU 基本権憲章、EU 条約、EU 機能条約を含む EU 法に明記されている EU の一般原則と価値観に沿ったものでなければならない。
- II. **AI 法及び国際的なアプローチとの整合性** - 施策、下位施策及び KPI は、AI 法の適切な適用に資するものでなければならない。これには、AI 法第 56 条第 1 項に従って、国際的なアプローチ（AI 安全機構又は標準設定機関が開発した標準又は指標を含む）を考慮に入れることが含まれる。
- III. すなわち、(a)所期の目的を達成するために適切であること、(b)所期の目的を達成するために必要であること、(c)達成しようとする目的に照らして過大な負担を課すべきではないこと、を意味する。比例性の具体的な適用例としては、以下のようなものがある：
 - a) 施策、下位施策及び KPI は、より重大なリスク又は重大な危害をもたらす不確実なリスクに対しては、より厳格であるべきである。規程は、例えば、深刻なリスクに関連する各サブメジャーに多数の KPI を設定することにより、汎用 AI モデルのプロバイダに対して、その深刻なリスクを緩和するための施策を講じること、又は深刻なリスクが発生する可能性が極めて稀であることを実証することを要求することができる。また、規程は、「ifthen」要件を用いるなどして、リスク軽減のための下位施策をリスクアセスメント KPI に関連付けることもできる。例えば、システミックリスクを有する汎用 AI モデルが能力 X を有すると評価された場合、Z の KPI を指針として、Y のリスク緩和を実施しなければならない。
 - b) 下位施策と KPI は具体的であるべきである。我々は、「対策」が「サブ対策」及び「サブ対策」への適合を証明する KPI よりも高い一般性で記述されることを認める。しかし、汎用の AI モデルプロバイダは、KPI によって証明される適切な「サブ対策」を満たす方法をできるだけ明確に理解すべきである。また、下位施策と KPI は、迂回や誤定義に対して堅牢でなければならない。本規定は、例えば、不必要な代理用語や指標の使用を避けることで、これを達成することができる。AI 事務局は、迂回やその他の誤指定の可能性のある下位施策と KPI を監視し、レビューする。
 - c) 施策、下位施策及び KPI は、該当する場合には、リスクの種類、配分戦略、展開の状況、及びリスクの程度に影響を及ぼす可能性のあるその他の要因を区別し、リスクをどのように評価し緩和する必要があるかを区別すべきである。例えば、システミックリスクの評価及び緩和に関するアセスメント、サブメント及び KPI は、意図的なリスクと非意図的なリスク（ミスアライメントを含む）を区別する必要があるかもしれないし、リスクの種類、流通戦略（オープンソース等）及び展開の状況によっては、他のリスクよりも具体的に厳格なものになるかもしれないし、そうでないかもしれない。
- IV. **将来性** - 下位施策と KPI は、AI 事務局が優れた情報に基づいて遵守の評価を改善する能力を維持すべきである。さらに、下位施策と KPI の更新プロセスは、急速な技術革新により、機動的な規制の策定と修正が必要となる可能性があることを想定すべきである。従って、具体的な要件と、技術や産業の発展に合わせて規則を適応・更新する柔軟性との間で、バランスを取る必要がある。規程は、例えばプロバイダが自ら監視し、検討することが期待できる動的な情報源を参照することで、これを達成することができる。そのような情報源の例としては、インシ

データデータベース、コンセンサス標準、リスク登録、リスクマネジメント枠組み、AI 事務局ガイダンスなどが考えられる。本規程は、サブ対策及び KPI が通常よりも迅速に更新されるよう努めるものとする。また、例えば、エージェント型 AI システムで使用されるモデルなど、新たな下位施策及び KPI を必要とするモデルの種類を明確にする必要があるかもしれない。

- V. **汎用 AI モデルのプロバイダの規模に対する比例性** - 汎用 AI モデルのプロバイダに適用される義務に関連する施策及び KPI は、汎用 AI モデルのプロバイダの規模を十分に考慮し、適切な場合には、AI 開発の最前線にある企業よりも資金力の乏しい中小企業や新興企業に対して簡素化された遵守方法を認めるべきである。システミック・リスクを伴う汎用 AI モデルのプロバイダに適用される義務に関連する KPI は、適切な場合には、プロバイダの規模や能力の違いも反映しなければならない。
- VI. **AI 安全エコシステムの支援と成長** - 我々は、汎用 AI モデルの開発、採用、ガバナンスが世界的な課題であることを認識している。本ドラフトにおける多くの方策は、例えば、モデルプロバイダ間で汎用 AI 安全インフラやベストプラクティスを共有するだけでなく、市民社会、学界、第三者、政府組織の貢献をさらに可能にするなど、異なるステークホルダー間の協力を可能にし、支援することを意図している。このため我々は、第 56 条 1 項 3 号および AI 法 116 条前文に沿って、関係者間の透明性を高め、知識を共有し、AI 安全性のための集成的で強固な証拠ベースを構築するために協力する努力を強化することを奨励する。また、オープンソースモデルが AI の安全エコシステムの発展にポジティブな影響を与えていることも認める。

目次

議長および副議長による序文	1
ドラフト計画と原則	3
目次	5
I. 前文.....	6
II. 汎用 AI モデルのプロバイダに対するルール	8
透明性	9
施策 1. AI オフィスの文書化.....	9
施策 2. 下流のプロバイダに対する文書化	9
附属書：利用規定に不可欠な要素	12
規則	13
施策 3. 著作権ポリシーを導入する	13
施策 4. TDM 例外の制限を遵守する。	13
施策 5. 透明性.....	14
III. システミック・リスクの分類法.....	16
施策 6. 分類	16
IV. システミック・リスクを伴う汎用 AI モデルのプロバイダに対するルール	19
施策 7. 安全・安心の枠組み.....	20
システミック・リスクを伴う汎用 AI モデルのプロバイダに対するリスクアセスメント	21
施策 8. リスクの特定	21
施策 9. リスク分析.....	21
施策 10. 証拠収集.....	22
施策 11. リスクアセスメントのライフサイクル	24
システミック・リスクを伴う汎用 AI モデルのプロバイダに対する技術的リスク緩和.....	25
施策 12. 緩和策.....	25
施策 13. 安全・安心レポート.....	26
施策 14. 開発と展開の決定	26
システミック・リスクを伴う汎用 AI モデルのプロバイダに対するガバナンス・リスク緩和.....	27
施策 15. システミック・リスク・オーナーシップ	27
施策 16. アドヒアランスとアセスメント	28
施策 17. 独立した専門家によるシステミック・リスクおよび緩和アセスメント	28
施策 18. 重大インシデント報告	29
施策 19 内部告発の保護	30
施策 20. 通知.....	30
施策 21. 文書化.....	31
施策 22. 公共の透明.....	32
結論	33

I. 前文

以下の通り

- a) 本実務規範の署名者は、域内市場の機能を改善し、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条件を整備する一方で、域内における AI の有害な影響に対して、健康、安全、民主主義、法の支配、環境保護を含む憲章に謳われている基本的権利の高水準の保護を確保し、AI 法（Act）第 1 条 1 項で強調されているイノベーションを支援することの重要性を認識する。本規定はこの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ック・リスクを伴う汎用 AI モデルのプロバイダに対して、AI 法第 53 条および第 55 条に定める義務の遵守に関するガイダンスを提供するものである。
- c) 本規定が汎用 AI モデルのプロバイダに言及する場合は常に、システムック・リスクを伴う汎用 AI モデルのプロバイダも包含するものとする。本規定がシステムック・リスクを有する汎用 AI モデルのプロバイダに言及する場合は常に、他の汎用 AI モデルのプロバイダを包含してはならない。
- d) 署名者は、本規定が GPAI モデルおよびシステムックリスクを伴う汎用 AI モデルのプロバイダが AI 法への準拠を証明する際の指針文書となることを認識すると同時に、本規定を遵守することが AI 法への準拠を証明する決定的な証拠とはならないことを認識する。
- e) 署名事業者は、AI 事務局と理事会が規程の適切性を定期的に監視・評価するために、規程の実施とその結果を報告することの重要性を認識する。
- f) 本規程は、AI 事務局による定期的な見直しを受けるものとする。AI 事務局は、AI 技術の進歩、社会の変化、新たなシステムック・リスクを反映させるため、本規程の更新を奨励・促進することができる。
- g) 署名者は、本規程が、汎用 AI モデルに関する統一 EU 標準が採用されるまでのつなぎとして機能することを認識している。将来の標準への段階的な移行を促進するために、更新が必要となる可能性がある。
- h) 署名者は、本規定に具体的な「対策」、「サブ対策」、「重要業績評価指標（KPI）」がないことは、システムックリスクを伴う汎用 AI モデルのプロバイダが、潜在的なシステムックリスクが顕在化した際に、それに対処し緩和する責任を免れるものではないことを認識する。
- i) AI 事務局と署名者は、汎用 AI モデルのプロバイダ、研究者、規制団体の連携を促進し、AI を取り巻く新たな課題と機会に対処するために協力する。

実務規範の目的は以下の通りである：

- I. 汎用 AI モデルのプロバイダは、効果的に義務を遵守することができる。実務規範は、プロバイダが遵守を証明する方法を明確にすべきである。また、同規範は、AI 事務局が、第 56 条に従って、遵守を証明するために同規範に依拠することを選択したプロバイダの遵守状況を評価できるようにすべきである。これには、汎用 AI モデル、特に最先端の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も含まれる。
- II. 汎用 AI モデルのプロバイダは、そのようなモデルを川下製品に統合できるようにするため、また AI 法やその他の規制（第 53 条および前文 101 参照）に基づくその後の義務を果たすために、AI のバリューチェーンに沿って汎用 AI モデルの十分な理解を効果的に確保することができる。
- III. 汎用 AI モデルのプロバイダは、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る（第 53 条および前文 106 参照）。
- IV. システミック・リスクを有する汎用 AI モデルのプロバイダは、システミック・リスクを有する汎用 AI モデルの開発、上市又は使用に起因する可能性のあるシステミック・リスクを、その発生源を含め、EU レベルで効果的に継続的に評価し、緩和することができる（第 55 条及びリサイタル 114 参照）。

II. 汎用 AI モデルのプロバイダに対するルール

以下の通り

- a) 署名者は、AI のバリューチェーンにおいて、汎用 AI モデルのプロバイダが果たす特定の役割と責任を認識する。なぜなら、プロバイダが提供するモデルは、様々な下流システムの基盤となる可能性があり、多くの場合、下流プロバイダは、そのようなモデルの製品への統合を可能にし、AI 法に基づく義務を果たすために、モデルとその機能について重要な理解を必要とするからである。¹。
- b) 署名者は、汎用の 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的 AI モデルは、アーティスト、作家、その他のクリエイター、およびクリエイティブなコンテンツの作成、配布、使用、消費の方法にとって、ユニークなイノベーションの機会であると同時に課題でもあることを認識する。さらに、著作権で保護されたコンテンツの利用には、関連する著作権の例外や制限が適用されない限り、関係する権利者の許諾が必要であることも認識している。²。
- c) 署名者は、モデルの修正や微調整の場合、プロバイダの義務は、比例性を守るために、修正や微調整に限定されるべきであると認識する。³。
- d) AI 法および本規定は、EU 法および国内法が定める規則を損なうものではなく、本規定は特に EU 著作権法に従って解釈されるものとする。指令 (EU) 2019/790 は、一定の条件の下で、テキストマイニングおよびデータマイニングを目的として、著作物またはその他の対象物の複製および抽出を認める例外と制限を導入した。この規則の下で、権利者は、科学的研究の目的で行われる場合を除き、テキストマイニングやデータマイニングを防止するために、著作物やその他の対象物に対する権利を留保することを選択することができる。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 モデルのプロバイダは、そのような著作物に対してテキストマイニングやデータマイニングを行いたい場合、権利者から許諾を得る必要がある。⁴
- e) 署名者は、AI 法第 53 条(1)(c)に従い、汎用の AI モデルを連邦市場に流通させるいかなるプロバイダーにも、著作権に関する連邦法を遵守するための方針を定める義務があることを認識する。53 (1) (c) AI 法によれば、汎用 AI モデルを上市するプロバイダは、著作権および関連する権利に関する連邦法を遵守するための方針を定める義務があり、特に、当該汎用 AI モデルの訓練を支える著作権関連行為が行われる法域にかかわらず、指令 (EU) 2019/790 の第 4 条 (3) に従い表明された権利留保を特定し、最先端技術を通じて遵守することが義務付けられている。⁵本セクションの第 2 章は、問題となっている様々な権利と正当な利益 () の間で公正なバランスを取りつつ、著作権の遵守と透明性を確保するための強固な枠組みを定めることによって、この義務 (⁶) の適切な適用に貢献することを目的としている。⁷これらの施策は、新興企業を含む中小企業 (SMEs) の利益にも十分配慮している。

従って、本規範の署名者は、以下を約束する：

¹前文 101。

²前文 105。

³前文 109。

⁴前文 105。

⁵前文 106。

⁶第 56 条 1 項

⁷CFEU 17 条 2 項、16 条、13 条、および 2008 年 1 月 29 日の CJEU 判決 Promusicae (C-275/06, ECR 2008 p. I271) ECLI:EU:C:2008:54, para 68 を参照のこと。ECLI:EU:C:2008:54, para 68; 2014 年 3 月 27 日判決、UPC Telekabel Wien (C-314/12) ECLI:EU:C:2014:192, para 46; 2022 年 4 月 26 日判決、ポーランド/議会と理事会 (C-401/19, Publié au Recueil numérique) ECLI:EU:C:2022:297, para 66

透明性

法律文書

第 53 条(1)(a)である：「汎用 AI モデルのプロバイダは、その訓練及び試験過程並びにその評価結果を含むモデルの 技術文書を作成し、常に最新の状態に維持しなければならない。

第 53 条(1)(b)である：「汎用 AI モデルのプロバイダは、その汎用 AI モデルを AI システムに統合しようとする AI システムのプロバイダに対して、情報および文書を作成し、最新の状態に保ち、利用可能にしなければならない。連邦法および国内法に従い、知的財産権、企業秘密または企業秘密を遵守し保護する必要性を損なうことなく、情報および文書は以下のものでなければならない：(i)AI システムのプロバイダが、汎用 AI モデルの能力と限界を十分に理解し、本規則に従った義務を遵守できるようにすること、(ii)最低限、附属書 XII に定める要素を含むこと。

施策 1. AI オフィスの文書化

署名者は、AI 事務局および各国所轄官庁の要請に応じて、下表に記載された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する。署名者は、透明性を高めるため、記載された情報の全部または一部を一般に開示できるかどうかを検討することが奨励される。

施策 2. 下流のプロバイダに対する文書化

署名者は、AI システムに汎用 AI モデルを組み込むことを意図する AI システムのプロバイダに対し、下表に示す情報および文書を作成し、常に最新の状態に保ち、利用可能にすることを約束する。署名者は、透明性を高めるため、記載された情報の全部または一部を公開できるかどうかを検討することが奨励される。

AI 法	情報の詳細	AI 事務局および各国所轄官庁向け	川下プロバイダ向け
附属書 XI §1 1.および附属書 XII 1.	一般情報：例えば、モデル名、バイナリを配布す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認証の証拠、開発者とモデルの所有者が同一でない場合はその法的商号、モデルファミリーの商号、提出された各モデルのバージョンの固有名などである。	✓	✓
附属書 XI §第 1 条 1 項(a)および附属書 XII 1 項(a)	意図されるタスク、および統合可能な AI システムの種類と性質：署名者は、意図する業務と、制限または禁止される業務の説明を提供する。この説明には、高リスクの AI アプリケーション（附属書 III に規定）がある場合、それを含め、汎用 AI モデルを組み込むことができる AI システムの種類と性質も記載する。	✓	✓
附属書 XI §1 1.(b)および附属書 XII 1.(b)を参照のこと。	利用規定：署名者は、プロバイダ間の共通慣行に基づく利用規定（AUP）の詳細を提供する。妥当性確認ポリシーは、最低限、附属書に定義された必須要素を含むものとする。署名者は、最新の利用規定の有効な URL を開示する。	✓	✓
附属書 XI §1 1.(c)および附属書	公開日と配布方法：公開日と配布方法：署名者は、汎用 AI モデルのすべての配布方法の最新リストとともに、公開日を提供する。署名者は、最	✓	✓

ドラフト・ドキュメント

AI 法	情報の詳細	AI 事務局および各国所轄官庁向け	川下プロバイダ向け
XII 1.(c)を参照のこと。	新の配布方法の有効な URL を開示する。		
附属書 XII 1.(d)	モデルと外部の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やソフトウェアとどのように相互作用するかについての文書を提供し、どのハードウェアとどのソフトウェアがモデルの一部でないかを明示する。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開示する。		✓
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 モデルの使用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。署名者は、必要なソフトウェアのバージョン依存関係を開示する。		✓
附属書 XI §1 1.(d)および附属書 XII 1.(f)を参照のこと。	アーキテクチャとパラメータ数：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総数、推論中に有効なパラメータ数についての説明を提供するものとする。	✓	✓
	署名者は、モデルのレイヤーの数や種類など、モデルのアーキテクチャについてより詳細に説明すべきである。	✓	
附属書 XI §第 1 条 1 項(e)、附属書 XII 1 項(g)および 2 項(b)	インプットとアウトプットの様式と形式：署名者は、該当する場合、インプットとアウトプットの様式、および関連する文脈上の制限を詳述する。	✓	✓
附属書 XI §1 1.(f)および附属書 XII 1.(h)を参照のこと。	ライセンス：署名者は、プロバイダ間の共通慣行に基づき、ライセンスの中核的要素を詳述する。これには、どのような資産（データ、モデルウェイトなど）が公開されるのか、また使用、変更、頒布の条件におけるライセンス義務に関する情報が含まれる。署名者は、最新のライセンスの有効な URL を開示する。	✓	✓
附属書 XI §第 1 条 2 項(a)および附属書 XII 2 項(a)	AI システムに統合するための技術的手段：署名者は、汎用 AI モデルを AI システムに適切に組み込むために必要な技術文書、インフラ、ツールを詳述する。署名者は、必要なソフトウェアやハードウェアのバージョン依存関係を開示する。	✓	✓
附属書 XI セクション 1 2.(b)	設計仕様と訓練プロセス：署名者は、モデル訓練の核となる要素（訓練段階、最適化の目的、最適化の方法、制約条件など）、設計決定に関連する根拠と仮定、その他訓練の詳細を詳述する。	✓	
附属書 XI §第 1 条 2 項(c)および附属書 XII 2 項(c)	訓練、テスト、妥当性確認に使用したデータに関する情報：データ取得方法、データ取得方法ごとの具体的な情報（ウェブクロール、データライセンス、データ注釈、合成データ、ユーザーデータなど）、データ処理に関する詳細（有害データや個人情報のフィルタリングの有無と方法など）、モデルの訓練/テスト/妥当性確認に使用したデータに関する具体的な情報（異	✓	✓

ドラフト・ドキュメント

AI 法	情報の詳細	AI 事務局および各国所轄官庁向け	川下プロバイダ向け
	なるデータソースからのデータの割合、訓練データ、テストデータ、妥当性確認データの主な特徴などを詳述する。		
	署名者はさらに、各データ形式（テキスト、画像、動画など）の訓練用データ、テスト用データ、妥当性確認データのサイズ（データポイント数）、データソースの不適性やデータのバイアスを検出するために使用した方法を詳述する。	✓	
附属書 XI セクション 1 2.(d)	計算資源：署名者は、AI 法第 97 条に従って採択された委任法に従い、比較可能で検証可能な文書を作成するための測定・計算方法を詳述するため、モデルの訓練と推論に使用される計算資源（例：汎用 AI モデルの訓練と推論に必要なハードウェアユニットの数と種類、訓練プロセスの期間、FLOP 数）を詳述する。	✓	
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、AI 法第 97 条に従って採択された、測定および算出方法を詳細に規定する委任法に従い、エネルギー消費量を評価するためにどのような情報および方法論（ハードウェアのプロバイダ、所在地、ハードウェアに関連するエネルギー源、消費されたエネルギー、発生した推定排出量など）を用いているかを詳細に規定する。	✓	
第 53 条 (1) (a)	試験過程とその結果：署名者は、試験を実施しない場合も含め、汎用 AI モデルについて実施する試験プロセスを詳述する。この詳細には、適切な解釈を確実にするため、実施した試験とその結果の説明を含める。	✓	

公開質問

上表の項目について、規程はどのように詳細を提供すべきか。

附属書：利用規定に不可欠な要素

AUP (Acceptable Use Policy) とは、あるサービスやテクノロジーをどのように利用すればよいかを概説する一連のルールと定義される。何が許容され、何が許容されない行為かについて、利用者にガイドラインを示す文書である。AUP は、加盟者の汎用 AI モデルの用途と機能を説明する資料と矛盾しないものとする。署名者は、自社の汎用 AI モデルに関連するすべての必要な情報を下流プロバイダと共有し、下流プロバイダが、AI システムの使用目的であるタスクやユースケースに適用される既存の規制を遵守できるようにする。

AUP は少なくとも、以下を含むべきである：

- AUP の存在理由を説明する目的声明；
- 適用範囲とは、その方針が誰に適用され、どのようなリソースを対象とするかを定義するものである；
- 主な用途と利用者；
- リスクの高い AI の用途を含め、許可される活動やタスクを列挙した許容される用途 (附属書 III に規定されている)がある場合、そのモデルは統合されることを意図している
- 許されない使用、禁止されている行為の詳細；
- 汎用 AI システムの利用者が従うべきセキュリティ・プロトコルの説明を含むセキュリティ対策；
- モニタリングとプライバシー、汎用 AI プロバイダがその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；
- AUP を遵守しなかった場合の警告プロセスと、ユーザー権限の停止または取り消しの規準；
- ユーザーアカウントを停止するための規準と、適用法および施行規則への言及；
- AUP を読み、理解し、遵守することに同意したことを認めるよう、ダウンストリーム・プロバイダに求める。

規則

法律文書

第 53 条(1)(c)である：「汎用 AI モデルのプロバイダは、著作権および関連する権利に関する連邦法を遵守し、特に、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、最先端技術を通じて特定し、遵守するための方針を定めなければならない。

施策 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連邦法を遵守するための方針を導入することを約束する。

施策 3 を達成するために、以下を行う：

下位施策 3.1. 著作権ポリシーを策定し、実施する

署名者は、当規範の本章に沿い、著作権および関連する権利に関する連邦法を遵守するための内部方針を策定し、実施する。この方針は、本章の対象となる汎用 AI モデルのライフサイクル全体⁸）を対象とする。汎用 AI モデルの修正または微調整の場合、汎用 AI モデルのプロバイダに対する義務は、AI 事務局が提供するガイダンスに従い、新たな学習データソースを含む修正または微調整にのみ及び⁹署名者は、本方針の実施と監督について、組織内で責任を割り当てる。

下位施策 3.2. 上流の著作権コンプライアンス

署名者は、汎用 AI モデルの開発におけるデータセットの使用について第三者と契約を結ぶ前に、妥当な著作権デューデリジェンスを実施する。特に署名者は、指令（EU）2019/790 の第 4 条 3 項に従って表明された権利の留保を、サードパーティがどのように識別し遵守したかについて、サードパーティに情報を求めることが推奨される。

下位施策 3.3. 川下の著作権コンプライアンス

署名者は、汎用 AI モデルが組み込まれた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緩和するため、合理的な下流の著作権対策を実施する。¹⁰下流の著作権に関する方針は、署名者が自社の汎用 AI モデルを自社の AI システムに垂直統合しているか、または、契約関係に基づいて汎用 AI モデルが他の事業体に提供されているかを考慮すべきである¹¹特に、署名者は、自社の汎用 AI モデルの過剰適合を回避し、また、別の事業体に対する汎用 AI モデルの契約上の提供の締結または妥当性の確認を、保護される著作物と同一または認識できるほど類似した出力が繰り返し生成されることを回避するために適切な施策を講じるといふ当該事業体の約束に依存させることが推奨される。本下位施策は中小企業には適用されない

施策 4. TDM 例外の制限を遵守する。

署名者が汎用 AI モデルの開発のために指令（EU）2019/790 の第 2 条（2）に従いテキストマイニングおよびデータマイニングに従事する場合、著作権で保護されたコンテンツへの合法的なアクセスを確保し、指令（EU）2019/790 の第 4 条（3）に従い表明された権利留保を特定し遵守することを約束する。

⁸前文 65.

⁹前文 109.

¹⁰第 3 条 1 項と第 3 条 63 項である。

¹¹前文 97 および第 3 条 68 項。

施策 4.を達成するために、以下を行う：

下位施策 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、それに従うクローラーのみを採用する。

下位施策 4.2. 見つけやすさに影響しない

Regulation (EU) 2022/2065 の第 3 条(j)に定義されるオンライン検索エンジンを提供する署名者、またはそのようなプロバイダを管理する署名者は、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、その検索エンジンにおけるコンテンツの見つけやすさに悪影響を与えないよう、適切な施策を講じる。

下位施策 4.3. その他の適切な手段に関する最善の努力

署名者は、広く利用されている業界標準に従い、オンラインで公に利用可能なコンテンツの場合、指令（EU）2019/790 の第 4 条 3 項に従い、ソースレベルおよび/または作品レベルで権利留保を表明するための、その他の適切な機械可読手段を特定し、遵守するために最善の努力を払う。特に、署名者は、総体レベルでの権利留保の表明を可能にする、広く採用されているツールを導入することが奨励される。

下位施策 4.4. 権利留保の標準の共同開発へのコミットメント

欧州委員会の招請を受け、署名者は、影響を受ける権利者の十分な代表者である事業者、および標準化団体などのその他の関係者と善意の協議を行い、指令（EU）2019/790 の第 4 条（3）に従い権利留保を表明し、当該権利留保を特定し遵守するための相互運用可能な機械可読標準を開発する。欧州委員会は、会議を招集し、議長を務め、関連する利害関係者および汎用 AI プロバイダと適宜、プロバイダ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表することができる。本下位施策は中小企業には適用されない。ただし、中小企業は自主的にこれらの協議に参加することができる。

下位施策 4.5. 海賊版サイトをクロールしない

署名者は、欧州委員会の「模倣品・海賊版監視リスト」に掲載されているウェブサイトを除くなど、海賊版ソースをクロール活動から除外するための合理的な手段を講じる。また、署名者は、自らが設立された法域の関連公的機関が公表した類似の除外リストに従うよう奨励される。

施策 5. 透明性

署名者は、著作権および関連する権利に関する連邦法を遵守するために採用する施策について、十分な透明性を確保することを約束する。

対策 5.を達成するために、以下を行う：

下位施策 5.1. 権利予約遵守に関する情報公開

署名者は、指令（EU）2019/790 の第 4 条(3)に従い表明された権利の留保を特定し、遵守するために採用する施策に関する適切な情報を、可能な限り多くの EU 市民が広く理解できる言語で公表する。当該情報は、各締約国のウェブサイト上で容易にアクセスできるものとし、常に最新の状態で保たれるものとする。

下位施策 5.2. クローラー名と robots.txt の機能

ドラフト・ドキュメント

前述の「下位施策」に従った情報には、最低限、本規程の対象となる汎用 AI モデルの開発に署名者が使用するすべてのクローラーの名称と、クローリング時を含む、関連する robots.txt の機能が含まれる。

下位施策 5.3. 窓口の一本化と苦情処理

署名者は、権利者が電子的手段によって直接かつ迅速にコミュニケーションできるよう、単一の窓口を指定することが奨励される。特に、団体管理団体を含む権利者及びその代表者が、汎用 AI モデルの開発のために、著作物その他の保護対象の利用に関する苦情を申し立てられるようにし、適切な苦情処理手続を実施することが奨励される。

下位施策 5.4. データソースと文書化

署名者が、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方針を定める義務を果たしているかどうかを、AI 事務局が¹² 監視できるようにするため、¹³ 署名者は、訓練、試験、妥当性確認のために使用されるデータソースに関する情報、および汎用 AI の開発のために保護されたコンテンツにアクセスし使用する権限に関する情報を作成し、常に最新の状態に保ち、AI 事務局の要請に応じて提供する。

¹²第 89 条 1 項

¹³前文 108 および第 53 条 1 項。

III. システミック・リスクの分類法

以下の通り：

- a) 署名者は、システミックリスクの分類法には、システミックリスクの種類、性質、発生源が含まれることを認識する。
- b) 署名者は、分類法が作成されたことを認識し、疑義がある場合は、AI 法第 3 条第 2 項に定義される各リスクの重大性と確率、および AI 法第 3 条第 65 項に定義されるシステミックリスクの定義に照らして解釈すべきである。
- c) 署名者は、システミックリスクの分類法が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、時とともに変更されることを認識する。
- d) 署名者は、セクション III とセクション IV が一般的に汎用 AI モデルについて言及しており、AI システムについては言及していないが、汎用 AI モデルが AI システムにどのように展開され得るかを考慮することで、多くの場合、リスクの特定、アセスメント、評価、緩和が最善であることを認識する。汎用 AI モデルのプロバイダが、システミック・リスクを有する汎用 AI モデルに基づく AI システムも開発・運用する場合には、これらのシステムを考慮に入れて、（安全・安心の枠組みに記載されているような）リスクアセスメントと緩和を行うことになる。

従って、本規範の署名者は、以下を約束

施策 6. 分類

署名者は、システミックリスクの評価と緩和の基礎として、このシステミックリスク分類法の要素を活用する。

6.1. システミック・リスクの種類

署名者は以下をシステミックリスクとして扱う：

- **サイバー攻撃**：脆弱性の発見や悪用など、攻撃的なサイバー能力に関するリスク。
- **化学、生物、放射線、核のリスク**：化学・生物・放射性・核兵器のリスク：デュアルユース科学は、とりわけ兵器の開発、設計、獲得、使用を通じて、化学・生物・放射性・核兵器による攻撃を可能にするリスクである。
- **コントロールの喪失**：強力な自律型汎用 AI モデルを制御できないことに関する問題。
- **AI の研究開発にモデルを自動利用する**：AI の開発ペースが大幅に向上し、システミック・リスクを伴う汎用 AI モデルの予測不可能な開発につながる可能性がある。
- **説得と操作**：選挙妨害、メディアに対する信頼の失墜、知識の均質化や過度な単純化など、民主主義的価値や人権に対するリスクを伴う大規模な説得や操作、大規模な偽情報や誤報の促進。
- **大規模な差別**：個人、地域、社会に対する大規模な違法差別。

例えば、重大事故、大規模なプライバシーの侵害や監視、また汎用 AI モデルが、公衆衛生、安全、民主的プロセス、公共・経済的安全保障、重要インフラ、基本的権利、環境資源、経済的安定性、人間の主体性、社会全体に大規模な悪影響を及ぼす可能性がある他の方法を考慮する。

公開質問

あるリスクがシステミック・リスクであるかどうかを定義する際に、考慮すべき関連事項や規準は何か。

これらの検討事項や規準に基づき、どのリスクを優先的にシステミックリスクの主要な分類に加えるべきか。

システミックリスクの分類法は、AI が生成的な児童性的虐待素材や非合意の親密な画像をどのように扱うべきか？

6.2. システミック・リスクの性質

システミック・リスクの性質とは、リスクのアセスメントや緩和方法に影響を与えるリスクの主要な性質を指す。署名者は、システミック・リスクの性質に特に関連する以下の側面と、各側面における事例を考慮するものとするが、これは網羅的なものでも、相互に排他的なものでもない：

- **起源モデル能力、モデル分布**
- **リスクを推進している行為者**：国家、グループ、個人、自律的 AI エージェント、なし（例：明確なアクターが特定できない）
- **意図的なもの**：意図的、非意図的（ミスアライメントを含む）
- **新規性**：前例がない、前例のない
- **確率-重大度比**：低影響高確率、高影響低確率、高期待影響
- **リスクが顕在化する速度**：徐々に、突然、継続的に変化する
- **リスクが顕在化している間のリスクの可視性**：あからさま（オープン）、隠密（隠されている）
- **出来事の経過**：直線的、再帰的（フィードバックループ）、複合的、連鎖的（連鎖反応）

6.3. システミック・リスクの源泉

リスク要因]または「リスク推進要因」とも呼ばれるリスク源とは、単独または複合的にリスク（モデル盗難や広範なサイバー脆弱性など）を引き起こす要素（事象、構成要素、行為者およびその意図や活動など）を指す。署名者は、システミックリスクの発生源として、特に以下のものを挙げる：

6.3.1. 危険なモデルの能力

これらは、システミック・リスクを引き起こす可能性のあるモデル能力である。署名者は、これらの能力の多くが有益な用途においても重要であることを認識している。これらには以下が含まれる：

- サイバー攻撃能力、化学・生物・放射性・核（CBRN）能力、武器獲得・拡散能力
- 自律性、拡張性、新しい仕事を学ぶ適応性
- 自己複製、自己改善、他のモデルを訓練する能力
- 説得、操作、ごまかし
- 長期的展望に立ったプランニング、予測、立案
- 状況認識

6.3.2. 危険なモデルの傾向

ドラフト・ドキュメント

これらは、システミック・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ある。以下のようなものがある：

- 人間の意図や価値観とのズレ
- 欺く傾向がある
- バイアス
- コンファビュレーション
- 信頼性と安全性に欠ける
- "目標追求"、目標修正への抵抗、"権力追求"
- 他の AI モデル／システムと "共謀" してそうしている。

6.3.3. モデルのアフォーダンスと社会技術的背景

これらは、モデルによってもたらされるシステミック・リスクに影響を及ぼす可能性のある、モデルの能力や傾向を超えた要因である。これらは、システミック・リスクを伴う汎用 AI モデルの具体的な入力、構成、文脈的要素を包含する。これらには以下が含まれる：

- ガードレール撤去の可能性
- ツールへのアクセス（他のモデルを含む）
- モダリティ（新規モダリティおよび複合モダリティを含む）
- リリースと流通戦略
- 人間の監督
- モデルの流出（モデルの漏洩／盗難など）
- ビジネスユーザー数とエンドユーザー数
- モデルを悪用する悪質業者の数、能力、意思を含む、攻守のバランス
- 社会的脆弱性または適応
- 説明不足や透明性の欠如
- 技術の成熟度（すなわち、ある技術が特定のアプリケーションのコンテキスト内でどの程度成熟しているか）。

データ、モデル、使用におけるフィードバックループ

IV. システミック・リスクを伴う汎用 AI モデルのプロバイダに対するルール

説明欄

施策、下位施策及び KPI は、相応のものでなければならない。特に、特定のプロバイダの規模や能力、特に AI 開発の最前線にいるプロバイダよりも資金力の乏しい中小企業や新興企業、また、適切な場合には、比例性の原則を反映し、利益とリスクの両方を考慮した様々な流通戦略（例えば、オープンソース）に合わせて調整する必要がある。

現在のドラフトは、システミック・リスクを有する汎用モデルとそのプロバイダの両方が少数であることを前提に書かれている。今後のドラフトでは、システミック・リスクが最も大きいモデルに焦点を絞ることを目的とした、より詳細な段階的施策の導入など、これらの数が増加した場合には、大幅な変更が必要となる可能性がある。

すぐ下の "whereas "の部分は、セクション IV の前文である。ここでは、ハイレベルの原則が、施策、下位施策、KPI の解釈を導く。

最後に、これは最初のドラフトに過ぎない。ご意見をお待ちしている。関連する未解決の質問を強調したが、ドラフトの他の部分についての意見も歓迎する。また、異なるビジネスモデルや展開戦略に対して、本施策をより適切なものにするだけでなく、より比例したものにする方法についての提案も歓迎する。

第 2 作業部会、3、4 の議長と副議長。

法律文書

第 55 条 1 項「第 53 条及び第 54 条に掲げる義務に加え、システミック・リスクを有する汎用 AI モデルのプロバイダは、以下の義務を負う：

- (a) システミック・リスクの特定と緩和を目的としたモデルの敵対的テストの実施と文書化を含め、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) システミック・リスクを有する汎用 AI モデルの開発、上市、使用から生じる可能性のあるシステミック・リスクを、その発生源を含め、連邦レベルでアセスメントし緩和する；
- (c) 重大なインシデントおよびそれに対処するための可能な是正施策に関する関連情報を把握し、文書化し、AI 事務局および必要に応じて各国の所轄官庁に不当な遅延なく報告すること；
- (d) システミック・リスクを伴う汎用 AI モデルと、その物理的インフラストラクチャー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。

以下の通り：

- a) 署名者は、システミックリスクを伴う汎用 AI モデルのプロバイダは、モデルのライフサイクル全体を通じて適切な施策を講じ、AI のバリューチェーンに沿った関連アクターと協力し、改善と新たな能力に照らして定期的の実務を更新することで、

ドラフト・ドキュメント

リスクマネジメントが将来にわたって確実に機能するよう、システムリスクを継続的にアセスメントし緩和すべきであることを認識する。¹⁴。

- b) 署名者は、システムリスクを伴う汎用 AI モデルが、(i) 実質的なシステムリスクをもたらす可能性が高い場合、(ii) 能力と影響が不確実な場合、(iii) 輸入事業者に関連する専門知識がない場合、詳細なリスクアセスメント、緩和策、文書化が特に重要であると認識する。逆に、新たな汎用 AI モデルが、重大なシステムリスクが顕在化することなく、既に安全に展開されているシステムリスクを有する汎用 AI モデルが示すような高い影響力を示すと信じるに足る十分な理由があり、適切な緩和策の実施が十分である場合には、より包括的な施策の必要性は低い。規模や能力の異なるプロバイダ間で利用可能なリソースが異なることを考慮し、また比例性の原則を認識するため、中小企業や新興企業向けの簡略化されたコンプライアンス方法が必要に応じて提供される。
- c) 重要な専門知識を有し、システムリスクのアセスメントと緩和を支援できる立場にある組織が幅広く存在することを、署名者は認識する。
- d) 多くのリスクアセスメント手法には、多大な労力とコストがかかることを署名者は認識している。例えば、評価、ベストプラクティス、インフラストラクチャーを共有すること、あるいは、適切な場合には、業界組織によって促進される可能性のある有資格のサードパーティ・プロバイダーと協働することなどによって、「負担を分かち合う」ことを互いに奨励している。
- e) 署名者は、システムリスクの効果的な評価と緩和に照らして、アセスメント、サブメカニズム、KPI に疑義がある場合は、それを解釈する。

従って、本規範の署名者は、以下を約束

施策 7. 安全・安心の枠組み

この枠組みは、システムリスクを伴う汎用 AI モデルから生じるシステムリスクを積極的に評価し、比例的に緩和するために遵守するリスクマネジメント方針を詳述したものでなければならない（第 55 条(1)参照）。SSF の包括性と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予想されるシステムリスクの重大性に比例すべきである。SSF の構成要素のうち、最初に必要とされるドラフトは、本セクションの残りの部分に概説されている。

¹⁴前文 114 AI 法（「システムリスクを伴う汎用 AI モデルのプロバイダは、説明責任やガバナンス・プロセスなどのリスクマネジメント・ポリシーを導入し、市販後モニタリングを実施し、モデルのライフサイクル全体を通じて適切な施策を講じ、AI のバリュー・チェーンに沿った関連アクターと協力するなどして、システムリスクを継続的にアセスメントし、緩和すべきである」）。

システムック・リスクを伴う汎用 AI モデルのプロバイダに対するリスクアセスメント

施策 8. リスクの特定

SSF の一環として、署名者は、システムック・リスクを伴う汎用 AI モデルに起因する可能性のあるシステムック・リスクを継続的かつ徹底的に特定することを約束する。

施策 8 を達成するために、以下を行う：

下位施策 8.1. リスクを判断する

署名者は、システムック・リスクを伴う汎用 AI モデルの開発、上市、利用が提案されている場合に、特に関連するシステムック・リスクを決定し、特定する。この目的のため、分類法（セクション III）に記載されたシステムック・リスクを使用し、分類法の他の要素を参照するだけでなく、追加リスクを検討することもできる。

施策 9. リスク分析

SSF の一環として、識別されたシステムック・リスクの経路を継続的かつ徹底的に分析する。

対策 9 を達成するために、以下を行う：

下位施策 9.1. 方法論

署名者は、システムック・リスクを伴う汎用 AI モデルの開発・展開が、特定されたシステムック・リスクを引き起こす可能性のある経路と、そのような経路を通じてそのようなリスクが顕在化する確率を特定するために、強固なリスク分析手法を用いる。

下位施策 9.2. システムック・リスク指標へのマッピング

システムック・リスクを伴う汎用 AI モデルについて、署名者は、潜在的に危険なモデル能力、傾向、その他特定されたシステムック・リスクへの道筋を可能にするリスク源を特定し、マッピングし、これらの要素ごとにシステムック・リスク指標を提供する。

下位施策 9.3. 厳しさの段階

システムック・リスクを伴う汎用 AI モデルについて、署名者は、特定された危険なモデル能力、危険なモデル傾向、その他のリスク源を、適切な保護施策がない場合、リスクレベルが耐えられないと判断される重大度の段階を含め、最低でも重大度の段階に分類する。

公開質問

その厳しさの段階はどうなるのか？すでに標準やコンセンサスが形成されつつあるのだろうか？

「重大性」は「重大性」のレベルを表現する最良の方法なのか、それとも確率と重大性の組み合わせとしてのリスクの定義に混乱を生じさせる可能性があるのか。

下位施策 9.4. リスクの予測

署名者は、下位施策 9.2 で言及されているシステミックリスク指標を発動するモデルの開発予定時期を、SSF のベストエフォートに含める。

施策 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 モデルがもたらす具体的なシステミック・リスクについて、継続的な証拠収集プロセスを実施する。予測からベスト・イン・クラス評価まで、様々な手法を活用し、これらのモデルの能力、傾向、その他の効果を調査する

施策 10 を達成するために、以下を行う：

下位施策 10.1. モデル診断的証拠

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、署名者は、文献調査、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（アルゴリズム効率、計算機使用量、エネルギー使用量など）、市民社会、学術界、その他関連する利害関係者を巻き込んだ参加型手法など、幅広い方法を用いて、モデルが示すシステミックリスクのモデル診断的証拠を収集する。また、スケーリングモデルによる能力向上を予測するスケーリング法則に取り組むこともある。このセクションのすべての証拠収集と同様、これは、適格なサードパーティと共同で、あるいはサードパーティに委託して行うことができる。

下位施策 10.2. クラス最高の評価

署名者は、システミックリスクを伴う汎用 AI モデルの能力と限界を適切に評価するため、クラス最高のアセスメントを実施する。この評価は、システミックリスクを持つ AI モデルのライフサイクル中、最も適切な時期に、様々な適切な方法論（例えば、Q&A セット、ベンチマーク、レッドチームやその他の敵対的サンプル、人間による高揚研究、モデル生物、シミュレーション、機密資料の代理評価など）を用いて実施し、関連リスクについて適格な評価者（内部または外部）によって行われるものとする。これらの評価の深さは、評価されるリスク及び当該モデルがどの程度のリスクを追加するかについての不確実性に比例するものとする（非常に類似したモデルの挙動に関する既存の知識は、例えば、必要な評価の深さを低減する可能性がある）。

公開質問

ある評価方法が特定のモデルやリスクに対して適切かどうか、評価が十分に徹底されているかどうかは、どのような要因によって決まるのだろうか？

下位施策 10.3. 科学的厳密性とその他の品質要素

署名者は、高い科学的厳密性を備えた評価の実施を確保する。特にシステミックリスクの重大度が高い階層については、適格なサードパーティによる主要結果の妥当性確認を通じて、さらなる厳密性を確保する（評価指標 17 参照）。署名者は、システミックリスクを持つ汎用 AI モデルを適切に評価するための十分な時間、モデルへのアクセス、計算予算など、厳密な科学的標準に沿った作業を行うために必要な支援を、社内外の評価者に提供する。

公開質問

高い科学的厳密性はどのように運用されるべきか。ゴールドスタンダードとはどのようなもので、どのような場合にそれを逸脱すべきか（例えば初期の探索的研究を行う場合）。

下位施策 10.4. 能力の引き出し

ドラフト・ドキュメント

署名者は、モデルの能力を十分に引き出し、能力を過小評価するリスクを最小限に抑えるため、クラス最高レベルの能力を引き出し（微調整、迅速なエンジニアリング、足場、計算およびエンジニアリング予算など）を行って評価を実施する。

下位施策 10.5. システムの一部としてのモデル

署名者は、システムック・リスクを伴う汎用 AI モデルの能力と限界を、そのモデルが使用されることが意図され、合理的に予見可能な将来の AI システムを代表する AI システムだけでなく、そのモデルがシステムック・リスクを引き起こす最大の可能性が明らかになった AI システムにおいても、アセスメントで評価できるようにする。

公開質問

システムックリスクを伴う汎用 AI モデルをオープンソースモデルとして、あるいは BtoB 顧客に提供している 署名者に対し、この下位施策はどのように促進されるか。

下位施策 10.6. 多様な評価と一般

署名者は、一般性を示すために、モデルの使用状況に合わせて評価を行う。例えば、多言語モデルの言語ベースの評価は、英語だけでなく、欧州の多様性を考慮した多言語評価にも焦点を当てる。

下位施策 10.7. 探索的作業

署名者は、システムック・リスクを伴う汎用モデルについて、（市民社会や学会の代表者を含む）適格な第三者によるオープンエンドのレッド・チーミングなど、相当量の調査作業を確実に実施する。つまり、すでに識別したリスクや能力に関する証拠収集だけにとどまらず、こうした手法を通じて新たなリスクや新たな能力を識別するよう努める。

下位施策 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、それに付随するベストプラクティスを、AI エコシステムの関係者が広く利用できるよう努める。特に特定された場合、商業的機密情報、公共の安全、拡散リスク、将来の評価の妥当性確認のため、署名者は情報の共有を制限することができる。

公開質問

現在 AI セーフティの最先端で活躍している研究チームに過度なプレッシャーを与えず、評価、ツール、ベストプラクティスの共有を促進するチャンネル、組織、方法はどのようなものがあるだろうか？

この施策は、こうしたツールや慣行をゼロから開発する能力はそれほど高くないかもしれないが、利用することはできるかもしれない新興企業や中小企業にとって特に有益なのだろうか？

下位施策 10.9. 共有する

署名者が評価結果を AI 事務局や一般市民と共有する場合、透明性が高く、比較しやすい形式で行う。実証的な結果の不確実性や、使用した手法の限界について、透明性をもって報告する。

施策 11. リスクアセスメントのライフサイクル

署名者は、システミックリスクを伴う汎用 AI モデルの開発・展開の全ライフサイクルにおいて、少なくとも本施策の下位施策に概説されている段階、および緩和策の実施前後（施策 12 に概説されている緩和策の有効性評価を含む）において、継続的にリスクをアセスメントし、証拠を収集することを約束する。

施策 11 を達成するために、以下を行う：

下位施策 11.1. 訓練前

システミックリスクを含む汎用 AI モデルの訓練実行を開始する前に、署名者は必要に応じて SSF を更新し、評価者（内部および外部）が、署名者の SSF コミットメントに沿って、証拠収集の準備が整っていることを確認する。

下位施策 11.2. 訓練中

署名、定期的な節目（一例として、有効計算量が 4 倍増加する毎）に証拠を収集し、リスクに見合った進捗状況の安全・セキュリティ報告書（SSR、施策 13 参照）を更新する。ここでいう訓練とは、「大規模なデータコーパスでの事前学習」のみを意味するものではなく、例えば、教師あり微調整、強化学習段階、またはモデルを改良する同様の方法も含むものとする。

下位施策 11.3. 展開中

システミック・リスクを伴う汎用 AI モデルの展開中、署名者は、リスクアセスメントを再検討し、特に関連する評価（および/または、より新しく改善された評価）を少なくとも 6 ヶ月ごとに再実行し、あるいは、（内部または外部）状況の大きな変化を認識した場合、または、その他の理由で以前のリスク評価結果を疑う理由がある場合はいつでも、展開中のモデルのモニタリングから得られた証拠も考慮に入れ、モデルの SSR を更新する。

下位施策 11.4. 展開後のモニタリング

署名者は、システミックリスクについて、展開後のモニタリングを実施する。また、展開後の関連情報を継続的に収集し、リスクアセスメントに含める仕組みを確立する。このような仕組みは、モデルの統合や利用方法によって異なる（例えば、有害な出力や行為に関するモニタリング、システミックな影響の調査など）。署名者は、配布戦略、モデルを使用する顧客や業界の種類に応じて、導入後のモニタリングを行う（例えば、オープンウェイトモデルの場合、ライセンスの遵守状況の評価、実社会におけるモデル使用の証拠の監視、モデルの科学的分析の調査など）。モデルプロバイダー自身が AI システムを展開する場合には、これらのシステムの一部としてモデルを監視する。

公開質問

システミック・リスクを考慮したオープンウェイトの汎用 AI モデルのプロバイダが、そのモデルの川下ユーザーに大きな副作用を与えることなく、公表したモデルを監視できるような方法は存在するか（あるいは存在しうるか）。

システミック・リスクを伴う汎用 AI モデルのプロバイダに対する技術的リスク緩和

施策 12. 緩和策

署名者は SSF の中で、システミックリスクの各指標または重大度の段階から、それに比例して必要とされる安全・セキュリティ緩和策までのマッピングを詳述し、利用可能な場合は AI オフィスのガイダンスを参考にすることを約束する。マッピングは最低限、システミックリスクを耐え難いレベル以下に抑えるよう設計され、それ以上のリスクをどのように最小化するかについても記述される。

施策 12.を達成するために、以下を行う：

下位施策 12.1. 安全緩和

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・リスクを緩和するために実施する安全緩和策を、SSF の中で詳述する。これらの安全緩和は、システミックリスクの指標や重大度に比例したものでなければならず、(a)モデルの挙動修正、(b)システム展開のためにモデルの周囲に設置されたセーフガード、(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ようにした対策やその他の安全ツールを伴う可能性がある。

下位施策 12.2. セキュリティ緩和

署名者は、(a)システミックリスクを持つ汎用 AI モデルの未公開ウェイト、(b)そのような未公開モデルを訓練または使用するために必要な、関連する未公開資産や情報の所有によるシステミック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。未公開モデルについては、展開決定を正当化するのに十分なリスクアセスメントが実施される前の開発ステージにおいて、これらのセキュリティ緩和を適用すべきである。リリースされたクローズドモデルについては、これらのセキュリティ緩和策をモデルの展開中及び展開後にも適用すべきであるが、このような緩和策は、ウェイト又は関連資産がオープンにリリースされたモデルには適用する必要はない。これらのセキュリティ緩和は、さらに、システミックリスク指標又は重大性の階層に比例すべきであり、(a)適切なハードウェアレベルを含む、静止時、移動時及び使用時の分銅及び資産の保護、(b)分銅及び資産へのアクセス管理、監視及び強化されたインターフェース、(c)継続的なセキュリティレッドチーム及び認定セキュリティレビューを通じた保証、並びに(d)内部脅威のスクリーニングを伴う可能性がある。

公開質問

システミック・リスクを伴う汎用 AI モデルには、システミック・リスクの指標や重大性の階層に応じて、どのようなサイバーセキュリティや情報セキュリティの標準を適用すべきか。

システミック・リスクを伴う汎用 AI モデルのサイバーセキュリティ標準は、他の領域における既存のサイバーセキュリティ標準とどのような点で異なるべきか。

下位施策 12.3. 制限事項

署名者は SSF の中で、既存の安全・セキュリティ緩和策の限界を詳述し、システミック・リスクの指標や重大度のレベルに応じて、システミック・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ことを明記する。

下位施策 12.4. マッピングの適切性を評価するプロセス

署名者は、下位施策 12.1.-12.2 における、システミックリスクの指標または重大度の階層から安全・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、SSF に詳述する。これは、施策 17.で概説した SSF 全体としての適切性を評価するための全体的なプロセスを超えて、能力引き出しの進歩やサイバーセキュリティの状況など、モデルの影響に関連する内的

及び外的要因の変化に対応するために実施されなければならない。

施策 13. 安全・安心レポート

リスク緩和とアセスメントの一環として、比較可能で検証可能な文書化を行う。

対策 8.-12., 署名者は、システミックリスクを伴う汎用 AI モデルを開発した場合、安全・セキュリティ報告書（SSR）を作成する。この報告書は、(a) モデル開発・展開のライフサイクルにおける適切な決定時点で作成され、(b) モデルのリスクと緩和アセスメントを詳述し、(c) モデルの開発・展開決定の基礎となる。

施策 13 を達成するために、以下を行う：

下位施策 13.1. 比例性

署名者は、SSR の(a)包括性と詳細度、(b)開発・展開ライフサイクルにおける時期、(c)外部からのインプットと精査のレベルが、すべてアセスメント対象のモデルに関連するシステミックリスク指標または重大度の階層に 比例していることを確認する。

下位施策 13.2. リスクアセスメントの結果

署名者は、緩和施策の実施前と実施後の両方で、当該モデルについて実施されたリスクアセスメントの結果を、対策 8.-11 に沿って SSR に詳述する。

下位施策 13.3. 安全緩和アセスメントの結果

署名者は、下位施策 12.1 に従い、実施された安全緩和施策の有効性のアセスメント結果を SSR に詳述する。

下位施策 13.4. セキュリティ緩和のアセスメント結果

署名者は、下位施策 12.2 に従い、実施されたセキュリティ緩和策の有効性のアセスメント結果を SSR に詳述する。

下位施策 13.5. 費用便益分析

署名者は、下位施策 14.2 に従った展開を正当化するための費用便益分析を SSR に詳述する。

下位施策 13.6. 方法論に関する十分な詳細

署名者は、下位施策 13.2.-13.5（下位施策 10.3 も参照）の結果、証拠、分析に使用された方法を独自に評価できるように、SSR に十分な科学的詳細が記載されていることを確認する。

下位施策 13.7. レビュー

署名者は、下位施策 13.2.-13.6 に記載された結果について、内部（または重大度が高い場合は外部）レビューの結果を SSR に記載する。

下位施策 13.8. 同等性

署名者は、AI 事務局と共有する 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する。

施策 14. 開発と展開の決定

安全・セキュリティ緩和が不十分であることによるリスクを軽減するため、署名者は、システミックリスクを伴う汎用 AI モデルの開発・

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。このプロセスは、署名者の SSF に記載され、SSR に提示された結果と分析に基づくものとする。

施策 14 を達成するために、以下を行う：

下位施策 14.1. 手続きを進めない条件

署名者は、SSF の中で、システミック・リスクを伴う汎用 AI モデルの開発・展開が進まない、あるいは、システミック・リスクを伴う既存の汎用 AI モデルが、安全・セキュリティ緩和が実施された後、そのモデルの SSR に基づき、展開から外されるか、削除される条件を詳述する。

下位施策 14.2. 手続きを進めるための条件

例えば、より優れた安全・セキュリティ緩和策の実施や、費用便益分析の提示などである。これらの条件については、システミック・リスクの指標や重大度に応じた厳密さとアセスメントプロセスを伴うものとする。

下位施策 14.3. 外部からの意見と意思決定

署名者は、開発および展開の決定において、AI 事務局などの政府関係者を含む外部関係者の意見を聞いたり、許可を得たりする必要がある場合、SSF にその旨を明記する。

システミック・リスクを伴う汎用 AI モデルのプロバイダに対するガバナンス・リスク緩和

施策 15. システミック・リスク・オーナーシップ

署名者は、システミック・リスクを評価し、比例的に緩和するために、経営陣や取締役会レベルを含むすべての組織レベルにおいて、システミック・リスクに関する適切なオーナーシップを確保することを約束する（第 55 条 1 項およびレシタル 114 参照）。55(1) および前文 114 参照）。

施策 15 を達成するために、以下を行う：

下位施策 15.1. エグゼクティブレベル

署名者は、システミック・リスクを伴う汎用 AI モデルが生み出すシステミック・リスクに対処するため、経営幹部レベルで責任とリソースを割り当てる。

下位施策 15.2. 取締役会レベル

署名者は、システミック・リスクを伴う汎用 AI モデルが生み出すシステミック・リスクについて、リスク委員会の設置など、取締役会レベル（またはそれに相当するレベル）で監督する責任とリソースを割り当てる。

公開質問

上記の下位施策は、プロバイダの規模やその他の関連する特性との関連で行うべきか。その場合、どのようにするか？

第 15 施策の遵守とみなされるような事例を、もっと、あるいはもっと増やすべきか？

施策 16. アドヒアランスとアセスメント

署名者は、自国の SSF の遵守と適切性を評価することを約束する（第 55 条 1 項および前文 114 項参照）。

施策 16 を達成するために、以下を行う：

下位施策 16.1. 定期的な SSF アセスメント

署名者は、計画されている活動を考慮し、SSF の適切性と順守状況について毎年アセスメントを実施し、文書化し、理事会またはそれに相当する機関に提出する。

公開質問

そのようなアセスメントが答えるべき特定の質問はあるのか？

この文脈において、妥当性はどのように定義されるべきなのだろうか？

施策 17. 独立した専門家によるシステムリスクおよび緩和アセスメント

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、特に重要度の高い階層において、適切な独立した専門家による有意義なリスクアセスメントと緩和策アセスメントを可能にする。このような独立した専門家によるリスク及び緩和の評価には、モデルの能力、収集された証拠、システムリスク、緩和の適切性の独立したテストが含まれる。また、独立した専門家による SSF および SSR のレビューを含むこともある（第 55 条 1 項および前文 114 参照）。

公開質問

システムリスクを有する汎用 AI モデルの独立した専門家によるシステムリスク・アセスメントは、どのような状況下で展開前に行うのが適切か？ 緩和策のアセスメントについてはどうか。どのような状況であれば、それは逆効果となるのか、あるいは不要となるのか？

リスクアセスメントに、独立した専門家を、訓練前または訓練中から、ライフサイクル全体を通じて、反復的に関与させることが適切または望ましい状況はあるか？

独立したシステムリスク・アセスメントは、関連するシステムリスクの大きさと性質、例えば、情報セキュリティ、システムリスクの要素を含む汎用 AI モデルや文書へのアクセスの深さ、テストの範囲、テストに要する時間、専門知識、透明性に関して、どのように適合させることができるか。

重大度レベルに対する対策はどうあるべきか？

施策 17 を達成するために、以下を行う：

下位施策 17.1. 展開前

リスクと緩和策をより正確に評価し、外部の関係者に保証を提供するため、システムリスクを伴う汎用 AI モデルを展開する前に、AI オフィスや適切なサードパーティ評価者などによる十分な独立した専門家によるテストを、利用可能な場合は AI オフィスのガイダンスに従って実施する。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。

公開質問

適切なサードパーティ評価者とは何か？業界の未成熟な現状を考慮し、どのように規程をドラフトすればよいか。リスクと緩和の独立した専門家による評価を確保するために、プロバイダ、特に中小企業を AI 事務局が支援する方法はあるか？

下位施策 17.2. 展開後

署名者は、システムック・リスクを伴う汎用 AI モデルについて、展開後の有意義な独立テストを適宜可能にし、例えばモデル・ライフサイクル全体のリスクアセスメントや、展開後の適切な修正を特定する。これには、独立した研究者だけでなく、AI 事務局を含むその他の関係者が、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、例えば、十分なアクセス、リソース、正当な研究活動に対する非報復の保証を提供する。

公開質問

独立したテストを容易にするためのさまざまな手段（リサーチのセーフハーバーや脆弱性報告など）は、どのような場合に適切なのだろうか？

施策 18. 重大インシデント報告

法律文書

第 55 条(1)(c)である：「第 53 条及び第 54 条に掲げる義務に加え、システムック・リスクを有する汎用 AI モデルのプロバイダは、重大なインシデント及びそれに対処するための可能な是正施策に関する関連情報を把握し、文書化し、過度の遅滞なく、AI 事務局及び必要に応じて各国の所轄当局に報告しなければならない。

署名者は、システムック・リスクを伴う汎用 AI モデルに起因する限りにおいて、重大インシデントを特定・追跡し、関連情報および可能な是正施策を文書化し、過度な遅延なく AI 事務局および必要に応じて各国所轄官庁に報告する。

公開質問

重大インシデントとは何を指すのか？規程は、AI 法が第 3 条 49 項で AI システムに用いている定義を用いるべきか、それともシステムック・リスクを伴う汎用 AI モデルには別の定義の方が適切か。

システムック・リスクを持つ汎用 AI モデルが、間接的に重大インシデントの発生につながったと判断されるには、どのような条件が必要なのだろうか。

AI 事務局への重大インシデント報告の自動化または合理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るか？

施策 18 を達成するために、以下を行う：

下位施策 18.1. 重大インシデント報告プロセス

署名者は、システムック・リスクを伴う汎用 AI モデルに起因する限りにおいて、重大インシデントおよびニアミスを特定し、文書化し、AI 事務局に報告するためのプロセス（職員の指名を含む）を確立する。

下位施策 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設定する。これには、重大なインシデントに対して講じられる可能性のある是正施策が、いつ講じられるかの説明とともにあらかじめ定義されていることが含まれる。

公開質問

重大なインシデントに対して、どのような是正施策が考えられるか。規範は、どのような場合に是正施策が適切であるかを明記すべきか。

オープンウェイトやオープンソースのプロバイダには、どのような重大なインシデント対応プロセスが適切なのだろうか？

施策 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当該違反の報告者の保護に適用される。"

署名者は、内部告発の手段を導入し、対象となる人物や活動に適切な内部告発保護を与えることを約束する。

施策 19 を達成するために、以下を行う：

下位施策 19.1. 情報を提供する

そのようなメールボックスが運用されている場合、署名者は、内部告発者の苦情を提出できるアル・オフィスのメールボックスを従業員に積極的に知らせる。

公開質問

EU 指令 2019/1937（「内部告発指令」）の他の部分で、規範で強調すべき重要な部分はあるか？

内部通報指令の中で、明確化すべき部分や、規程にさらに明記すべき部分はあるか？ システムック・リスクのアセスメントと緩和を可能にするために適切と思われる追加的な内部 通報施策はあるか？

施策 20. 通知

署名者は、システムック・リスクを有する汎用 AI モデルとして分類するための閾値を満たすモデル、その SSF、SSR、および必要に応じて実質的なシステムック・リスクに関する関連情報を AI 事務局に通知することを約束する。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 事務局の義務を理解した上で行われる。

施策 20 を達成するために、以下を行う：

下位施策 20.1. システムック・リスクを通知する汎用 AI モデル

法律文書

ドラフト・ドキュメント

第 52 条 1 項「汎用 AI モデルが第 51 条(1)の(a)の条件を満たす場合、当該プロバイダは、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、遅滞なく、いかなる場合においても、2 週間以内に欧州委員会に通知しなければならない。この通知には、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。欧州委員会は、通知されていない汎用 AI モデルがシステミック・リスクを有していることを知った場合、そのモデルをシステミック・リスクを有する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、訓練を開始する前に、使用する予定の計算能力を見積もり、その結果、汎用 AI モデルがシステミック・リスクを伴う汎用 AI に分類される場合は、AI 事務局に通知する。

公開質問

AI 事務局は、汎用モデルが高い影響力を有すると推定されるかどうか（したがって、システミック・リスクを有する汎用 AI モデルとして分類されるかどうか）を判断するための分類規準を更新する認可を有している。プロバイダが新たな分類基準に合致するモデルを AI 事務局に通知すべき時期が明確になるような書き方はどうか。

下位施策 20.2. SSF の通知

署名者は、AI 事務局が最新版の「安全・安心の枠組み」にアクセスできるようにする。

公開質問

このアクセスはどうすれば容易になるのか？

下位施策 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 モデルを上市する前に、その決定に先立ち AI 事務局に SSR を送付する。

下位施策 20.4. 実質的システミック・リスクの届出

署名者は、実質的なシステミック・リスクが顕在化する可能性があるかと確信する強い理由がある場合、AI 事務局に通知する。

公開質問

システミック・リスクが顕在化するかもしれないと考える強い理由とは何か？

施策 21. 文書化

署名者は、システミック・リスクを伴う汎用 AI モデルのライフサイクルを通じて、本規定および AI 法のシステミック・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、この情報を要請に応じて AI 事務局と共有することを約束する。

これには、附属書 XIII の情報など、システミック・リスクを有する汎用 AI モデルの分類に関連する証拠が含まれる。また、附属書 XI 第 2 節（第 53 条(1)(a)参照）に概説されている情報に加え、SSF、SSR、リスクアセスメント中に収集された追加的な証拠など、AI 法および規範への準拠を立証する文書も含まれる。

公開質問

特に小規模プロバイダ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り得るか。注：今後のドラフトでは、この施策に基づく文書が合理化され、附属書 XI、セクション 1、附属書 XII に詳述されているような他の文書要件が組み合わされるようにする予定である。

施策 22. 公共の透明

署名者は、SSFとSSRを公表することにより、特にAIのリスクアセスメントとリスク緩和に関する科学が発展途上であることを踏まえ、システムック・リスクを伴う汎用AIから生じるシステムック・リスクを、川下プロバイダ、AIオフィス、一般市民を含むより広範なエコシステムがよりよく理解し緩和できるよう、適切な透明性を提供することを約束する。情報が含まれることでシステムック・リスクが大幅に増大する場合や、社会的利益に不釣り合いな程度に機密性の高い商業情報が漏洩する場合は、情報を編集することができる。

公開質問

システムック・リスクを評価し緩和するために、より広範なエコシステムに権限を与えることによって、システムック・リスクが減少するのではなく、どのような種類やレベルの公的透明性が増加するのだろうか？

モデルカードやシステムカードを公表する一般的な慣行を考えると、このような公的透明性はどの程度負担になるのか。このような負担を軽減するような施策は可能か？

結論

本規範の初期ドラフトは、4つの専門作業部会による既存のベストプラクティスの予備的レビュー、430件近い提出物によるステークホルダー協議のインプット、プロバイダ・ワークショップからの回答、国際的アプローチ（G7実務規範、フロンティアAI安全約束、ブレッチリー宣言、関連する政府および標準設定団体からのアウトプットを含む）、そして最も重要なこととしてAI法そのものを検討した結果である。この初期段階では、ドラフトは必然的にハイレベルなものとなり、主に実務規範の基礎となる原則と、いくつかの「施策」と「小施策」の案が示されている。

我々は、このドラフトがあくまで第1ドラフトであり、その結果、本規定案の提案は**暫定的なものであり、変更される可能性がある**ことを強調する。従って、本規程の内容をさらに発展・更新し、2025年5月1日に向けてより詳細な最終形を作成するため、皆様からの建設的なご意見をお聞かせいただきたい。特に以下の点にご留意いただきたい：

1. この第1ドラフトは、ドラフト計画で定められた**6つの重要な考慮事項**、すなわち、i) 組合の原則および価値観との整合性、ii) AI法および国際的アプローチとの整合性、iii) リスクとの比例性、iv) プロバイダの規模および能力との比例性、v) AIセーフティ・エコシステムの支援と成長、vi) 将来性、によって導かれる。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る。
2. 私たちは、市民社会、学術界、AI安全機構、産業界を含む様々な立場のアクターからのインプットに基づき、施策、下位施策、KPIを**包括的に見直し、発展させ、洗練させる必要性**を認識している。将来的には、前述のドラフト計画と原則に基づき、AI法の条文やリサイタルへの具体的な言及を含むことになる。上記の例は予備的なものであるが、今後の策定において、「対策」、「サブ対策」、「KPI」について、詳細なコメントを歓迎する。また、「施策」、「下位施策」、「KPI」をどのようにすれば、より適切なものとし、異なるビジネスモデルや展開戦略に対してより適切なものとすることができるかについての提案や、ドラフトに記載されている未解決の問題をどのように解決するかについての提案も歓迎する。
3. 現在のドラフトは、**システムック・リスクを有する汎用AIモデルも、そのプロバイダも少数しか存在しないという前提**で書かれていることに留意したい。その前提が誤りであることが判明した場合、将来のドラフトは、例えば、最大のシステムック・リスクをもたらすモデルに主眼を置くことを目的とした、より詳細な段階的施策システムを導入するなど、大幅に変更する必要があるかもしれない。

11月28日（木）12:00（中央ヨーロッパ標準時）までに、専用プラットフォーム（Futurium）上のフォームより、本ドラフトに関するご意見をお寄せいただきたい。私たちは、この規約の今後の改訂版について、皆様と協力できることを楽しみにしている。