汎用 AI 実務規範第 2 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 2 次ドラフトをここに提示する。各作業部会の参加者および実務規範総会のオブザーバーは、2025 年 1 月 15 日までに、このドラフトに対する意見書を、共有された専用アンケートを通じて提出することができる。

我々は、第 2 ドラフトが依然として作業途中であることを強調したい。第 1 ドラフトに対するフィードバックを受け取ってから、この第二ドラフトを公表するまでの期間が短かったため、私たちは主に、明確な説明を提供し、必要不可欠な詳細を追加し、比例性へのアプローチを洗練させることに重点を置いてきた。この段階で具体的なフィードバックを取り入れたとしても、最終的な規程に盛り込まれることを保証するものではない。逆に、本ドラフトにおいて特定の要素が変更されない場合、これは永続性を示すものではない。

本規範の第 2 ドラフトでは、4 つの作業部会が緊密に連携し、AI 法第 5 章に準拠する際の、汎用 AI モデルのプロバイダおよびシステミック・リスクを有する汎用 AI モデルのプロバイダにとっての重要な考慮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：システミックリスクのリスクアセスメント
- 第 3 作業部会：システミックリスクの技術的リスク緩和
- 第 4 作業部会：システミック・リスクのためのガバナンス・リスク緩和

第 1 作業部会 透明性は、AI 法第 53 条第 2 項に規定される条件を満たす無償のオープンソースライセンスの下で公開され、システミックリスクを伴う汎用 AI モデルに分類されないものを除き、すべての汎用 AI モデルに適用される。第 1 作業部会の著作権は、すべての汎用 AI モデルに適用される。第 2 作業部会、第 3 作業部会及び第 4 作業部会は、対応する第 3 章とともに、AI 法第 51 条に基づきシステミック・リスクを伴う汎用 AI モデルに分類される汎用 AI モデルのプロバイダにのみ適用される。

最初のドラフトに対してステークホルダーから寄せられたフィードバックを徹底的に検討した結果、私たちは実務規範の「目的」を維持しつつ、「約束」と「施策」を洗練させ、規定を拡大した。この第 2 ドラフトは、さらなる発展のための基礎となるものである。次回のドラフトでは、EU アンケート、プロバイダ・ワークショップ、作業部会の会合で提供されたご意見を参考にする予定である。これまでのところ、皆様からのフィードバックは非常に有益であり、大幅な変更が加えられている。**従って、関係者の皆様には、新しい要素も変更されない要素も含め、規程のあらゆる側面について包括的なフィードバックを提供し続けるようお願いしたい。**皆様からのフィードバックは、汎用 AI モデルの開発と展開の未来を導く上で重要な役割を果たす規範の最終版を形作る助けとなるだろう。

今回もハイレベルのドラフト計画を盛り込み、実務規範の指針となる原則とその前提を概説した。具体的なコミットメント、施策、主要業績評価指標（KPI）については、引き続き徹底的な審議を行うが、ドラフトプランにより、最終的な規程の形式と内容がステークホルダーにとって明確になることを期待している。

本規範の本バージョンに含まれる例示的な KPI は、暫定的なものであり、見直しと改訂の可能性がある ことに留意されたい。例えば、定量的なものもあれば、定性的なものもある。**従って、KPI に関するフィードバックを強く推奨し、歓迎する。**

AI 法は 2024 年 8 月 1 日に施行され、規程の最終版は 2025 年 5 月 2 日までに完成させる必要があるとされた。第 2 ドラフトは、これまでの作業を踏まえつつ、2025 年以降に開発・発売される次世代モデルに適した「将来性のある」規範を提供することを目的としている。

本第 2 ドラフトを策定するにあたり、本規程の適用範囲内の事項については、主に AI 法の規定に従った。従って、本規程の文脈や定義に別段の定めがない限り、本規程で使用されている用語は、AI 法の用語と同一である。本第 2 ドラフトでは、AI 法の規定を網羅的に参照していないが、将来的にはそうする予定である。

最初のドラフトと同様、この文書も、産・学・市民社会から数百人が参加した共同作業の結果である。第 1 ドラフトに対して寄せられたフィードバックは、我々のドラフト作成プロセスにおいて洞察に富み、有益なものであった。我々は、AI ガバナンス、国際的アプローチ（AI 法第 56 条 1 項に規定）、連邦法の慣行規範（偽情報に関する慣行規範など）、業界のベストプラクティス、プロバイダや作業部会メンバーの専門知識と経験について、進化する文献から引き続き情報を得ている。

規範の策定プロセスの主な特徴は以下の通りである：

- 専門知識、経験、独立性（金銭的利害関係がないことを含む）、性別や地理的な多様性を確保するために AI 事務局によって選出された委員長および副委員長によってドラフトされる。
- 9 月に締め切られたマルチステークホルダー協議には 427 件の応募があった。
- 規範の第 1 ドラフトに関するマルチステークホルダー調査には 354 件の応募があった。
- AI オフィスによる組織的サポートと法的アドバイス
- 議長と副議長によるプロバイダ・ワークショップ
- 議長と副議長が率いる 4 つの専門作業部会会議
- AI 理事会および欧州議会の EU 署名者代表者との会合

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。独立した委員長と副委員長のグループとして、私たちは、このプロセスをできるだけ透明性のあるものにし、利害関係者がアクセスしやすいものにするよう努力する。皆様の継続的な協力と建設的な批判を期待している。

最後に、ドラフト作成プロセスの現段階において、私たちが最優先していることのひとつは、 取り組んでいる問題に関する私たちの動機と理由を明確に伝えることである。これらの問題の多くは微妙で複雑なものであり、私たちは規範のドラフトを通じてそれらを透明性をもって伝えることを目指している。さらに、私たちは、実務規範のすべての部分がシームレスにまとまり、理解しやすくなるようにするよりも、主にコミットメント、施策、KPI を洗練させることに時間を使った。私たちは、実務規範を強化するため、以後の反復作業でこれらの点の改善に努める。

我々は、2025 年 1 月 15 日までに、実務規範総会参加者とオブザーバーが共有する専用アンケートを通じて、書面によるフィードバックを歓迎する。

応援ありがとう！

アレクサンダー **マティアス**

オリバー **ペウカート** **ベンジオ** **シャケ**

作業部会 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

共同議長

リシ **セリーヌ・カステッツ**

ボマサニ **レナール** **ジオシ** **プリヴィテラ** **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 **マーカス**

ザッヘル **ラージクマール** **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、原則、前提条件

この第 2 ドラフトは、より詳細な規定と具体例を示している。現段階では、特に KPI については、規程の最終的な採択版に盛り込まれることを期待され ているような詳細なレベルには至っていない。これは、(i) 規範の構成と原則について幅広い合意を得るための作業を行っている段階であること、(ii) より詳細な提案書を作成するための時間が十分でないこと、(iii) AI の最新の発展と進歩を反映させるため、継続的に規程の詳細を更新する予定であること、が理由である。本規範に記載されているコミットメントは、コミットメント、対策、KPI の下位階層で構成されている。これらの要素のうち、特に KPI がない場合、これは決定的な決定ではなく、この第 2 ドラフトの作成中に遭遇した時間的制約の結果である。さらに、このドラフトには、規範の見直しや更新方法に関する項目がまだ含まれていない。

以下は、私たちがドラフトを作成する際に遵守している基本原則である：

1. **EU の原則と価値観との整合**-コミットメント、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とする。

2. **AI 法及び国際的アプローチとの整合性** - コミットメント、施策及び KPI は、AI 法の適切な適用に貢献する。これには、AI 法第 56 条第 1 項に従い、国際的なアプローチ（AI 安全機構や標準設定機関が策定した標準や指標を含む）を考慮することが含まれる。

3. すなわち、(i)望ましい目的を達成するために適切であること、(ii)望ましい目的を 達成するために必要であること、(iii)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. コミットメント、施策、及び KPI は、より高いリスク階層又は深刻な危害の不確実なリスクに対して、 より厳格であるべきである。例えば、深刻なリスクに関連する「対策」ごとに複数の KPI を提示することで、本規定はこれを達成することができる。これにより、汎用 AI モデルのプロバイダは、深刻なリスクを緩和するための行動をとるか、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確 実に証明することが求められる。規程はまた、「if-then」要件を用いるなどして、リスク軽減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持つ汎用 AI モデルの能力が X と評価された場合、署名者は Z の KPI で追跡される Y のリ スク緩和策を講じることを約束する。

    b. 対策と KPI は具体的であるべきである。コミットメントはより一般性の高いレベルで明確化されるかもしれないが、汎用の AI モデルプロバイダは、適切な KPI によって追跡される「対策」をどのように満たすかについて明確な理解を持つべきである。評価指標及び KPI は、効果的であり、誤設定や回避の試みに対して堅牢であるように設計されるべきである。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により、これを達成するよう努める。AI 事務局は、迂回やその他の形の誤指定の可能性がある指標と KPI を監視し、見直す。

    c. コミットメント、対策、及び KPI は、該当する場合には、異なるタイプのリスク、当該汎用 AI モデルの流通戦略及び展開状況、並びにリスクの階層に影響を及ぼす可能性のあるその他の要因、及び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る必要がある。例えば、システミックリスクを評価し緩和するコミットメント、リスクアセスメント及び KPI は、意図的なリスクと非意図的なリスク（不整合の事例を含む）を区別する必要があるかもしれない。さらに、モデルウェイトが自由に公表されている場合、プロバイダがシステミックリ スクを評価

し緩和するために利用できる様々な手段を考慮するために、コミットメント を適合させる必要があるかもしれない。

4. **将来性** - AI 技術は急速に変化している。対策と KPI は AI を維持する必要がある。

   事務局は、新たな情報に基づき、遵守状況の評価を改善することができる。従って、 規程は、必要に応じて、「施策」と KPI の迅速な更新を促進するよう努めなければならない。一方では具体的な要求事項や実績指標を、他方では技術や業界の発展に迅速に適応できる柔軟性を、バランスよく見出すことが重要である。規程は、例えば、プロバイダがリスクアセスメントと緩和において監視し、考慮することが期待される動的な情報源を参照することによって、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、インシデントデータベース、コンセンサス標準、最新のリスク登録、最新のリスクマネジメント枠組み、AI 事務局のガイダンスなどが考えられる。技術が発展するにつれて、特定の汎用 AI モデル、例えば、エージェント型 AI システムで使用されるモデルについて、追加的な対策と KPI を明示することも必要になるかもしれない。

5. **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 開発の最前線にいる企業よりも資金力の乏しい中小企業（SME）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。

6. **安全で、人間中心で、信頼できる AI のための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、ガバナンスがグローバルな問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ガバナンスは、例えば、モデルプロバイダー間で汎用 AI の安全インフラやベストプラクティスを共有したり、証拠収集において市民社会、学術界、サードパーティ、政府組織の参加を奨励するなど、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。我々は、第 56 条（1）（3）、前文 1、前文 116 に則り、安全で人間中心の、 信頼できる AI のための集合的で強固な証拠ベースを構築するために、利害関係者間の透明性を高め、知識を共有し、協力する努力を促進する。また、オープンソースモデルが、安全で人間中心かつ信頼できる AI の開発にポジティブな影響を与えていることも認める。

7. **AI ガバナンスとリスクマネジメントの革新** - 汎用 AI モデルの安全性を理解し確保するための最も効果的な方法を決定することは、依然として発展途上の課題であることを政府は認識している。当規範は、プロバイダが AI の安全ガバナンスと関連する証拠収集の方法と実践において、最先端の技術を競い合い、進歩させることを奨励すべきである。プロバイダが、より負担の少ない代替的なアプローチによって、同等かそれ以上の安全性成果を実証できた場合、こうしたイノベーションは、AI ガバナンスと証拠の最先端技術の改善として認められるべきであり、我々はその普及を支援すべきである。

現在のドラフトは、**システミックリスクを有する汎用 AI モデルも、そのプロバイダも少数であろうという前提で**書かれている。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、最大または最も深刻なシステミック・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、より詳細な段階的コミットメント・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。特に強調したいのは、**汎用 AI モデルの修正によって対象プロバイダが増加したとしても、AI 法第 53 条および第 55 条に基づく修正者の義務は、それぞれの修正範囲に限定されるべきであるという**ことである。これらの点については、専用の Q&A に記載されているように、AI 事務局からのさらなる明確化を期待したい

目次

議長および副議長による開会宣言 1

ドラフト計画、原則、前提条件 4

目次 7

I.前文 9

II.汎用 AI モデルのプロバイダによるコミットメント 11

透明性 11

コミットメント 1.文書化 12

附属書利用規定に不可欠な要素 20

著作権 21

コミットメント 2.著作権ポリシー 22

III.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よるコミットメント 26

コミットメント 3.分類法 28

コミットメント 4.安全・安心の枠組み 32

コミットメント 5.安全・安心モデルレポート 34

コミットメント 6.モデルのライフサイクルに沿ったリスクアセスメントと緩和 36

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 37

コミットメント 7.リスクの特定 37

コミットメント 8.リスク分析 38

コミットメント 9.リスク評価 38

コミットメント 10.証拠収集 39

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 45

コミットメント 11.安全緩和 45

コミットメント 12.セキュリティ緩和 46

コミットメント 13.開発と展開の決定 49

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 51

コミットメント 14.システミック・リスクの責任分担 51

コミットメント 15.枠組みの順守とアセスメント 52

コミットメント 16.外部リスクアセスメント 53

コミットメント 17.重大インシデントの報告 56

コミットメント 18 内部告発の保護 58

コミットメント 19.通知 60

コミットメント 20.文書化 61

コミットメント 21 公共の透明性 62

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た基本的権利の高水準の保護を確保し、AI 法第 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、システミック・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。本規定がシステミック・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。

c) 署名者は、本規定が、汎用 AI モデルおよびシステミック・リスクを伴う汎用 AI モデルのプロバイダが AI 法を遵守していることを証明する際の指針となる文書であることを認識する一方、本規定を遵守することが AI 法遵守の決定的な証拠となるわけではないことを認識する。

d) 署名者は、AI 事務局と理事会による規程の適切性の定期的な監視と評価を促進するため、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（AI 法第 56 条 5 項）。

e) 署名者は、本規程が AI 事務局による定期的な見直しの対象となることを認識する。AI 事務局は、AI 技術の進歩、社会の変化、新たなシステミック・リスクを反映するため、本規程の更新を奨励・促進することができる（AI 法第 56 条 6 項）。

f) 署名者は、本規程が、汎用 AI モデルの標準規格が EU で統一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、更新が必要な場合もある。

g) 署名者は、本規定に具体的なコミットメント、対策、重要業績評価指標（KPI）が存在しないことは、システミックリスクを伴う汎用 AI モデルのプロバイダが、潜在的なシステミ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。

h) 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 を取り巻く環境における新たな課題と機会に対処するため、AI 事務局とのパートナーシップの重要性を認識している。

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、AI 法に基づく義務を効果的に遵守することができる。実務規範は、プロバイダに対して、どのように遵守を証明するかを明確にすべきである。また、AI 事務局は、AI 法第 53 条第 4 項および第 55 条第 2 項に従って、遵守を証明するために本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（AI 法第 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（AI 法第 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、連合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（AI 法第 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よるコミットメント

以下の通り

a) なぜなら、提供されるモデルは、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しているからである（AI 法 101 項参照）。

b) 署名者は、モデルの修正や微調整の場合、比例性を守るため、プロバイダの義務は修正や微調整に限定されるべきであると認識する（AI 法 109 項参照）。

c) 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。権利の留保が適切な方法で表明されている場合、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（AI 法 105 条参照）。

従って、本規範の署名者は、以下を約束する：

透明性

法律文書
AI 法第 53 条第 1 項（a）："汎用 AI モデルのプロバイダは，AI 事務局及び国内所轄官庁に要求に応じて提供する目的で，最低限附属書 XI に定める情報を含む，訓練及び試験の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，かつ，最新の状態に維持しなければならない。" AI 法第 53 条第 1 項（b）：「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。" AI 法第 53 条第 2 項「第 1 項の(a)及び(b)に定める義務は、モデルへのアクセス、利用、修正及び頒布を可能とする無償のオープンソースライセンスの下でリリースされ、重み、モデル構造に関する情報及びモデルの利用に関する情報を含むパラメータが一般に公開されている AI モデルのプロバイダには適用されない。この例外は、システミック・リスクを有する汎用 AI モデルには適用されない。"

AI 法第 53 条第 7 項「営業秘密を含め、本条に従って入手した情報または文書は、第 78 条に定める守秘義務に従って取り扱われるものとする。

コミットメント 1. 文書化

AI 法第 53 条 1 項(a)および(b)の義務を履行するため、署名者は以下の施策を講じる。本施策は、AI 法第 53 条第 2 項に規定される条件を満たすオープンソース AI モデルのプロバイダには適用されないが、そのモデルがシステミックリスクを有する汎用 AI モデルである場合はこの限りではない。システミック・リスクを伴う汎用 AI モデルのプロバイダである署名者には、本規程の第 20.2 項が適用され、AI 法第 53 条第 1 項(a)号が要求する追加書類（具体的には、AI 法附属書 XI 第 2 項に記載される書類）が対象となる。

施策 1.1. 関連情報を作成し、最新の状態に保ち、プロバイダに提供する。

署名者は、AI 法第 53 条(1)、(b)および(7)に規定される営業秘密および機密保持の保護を受けて、以下の表 1 に記載される情報を作成し、要請に応じて AI 事務局および各国所轄官庁、および/または川下プロバイダに提供することを約束する。

署名者は、透明性を促進するため、文書化された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奨励される。また、これらの情報の一部は、AI 事務局が提供するテンプレートに基づき決定される、AI 法第 53 条 1 項（d）に基づきプロバイダが公開しなければならない訓練内容の要約の一部として、要約された形で要求される場合もある。

署名者は、汎用 AI モデルの変更を反映させることも含め、文書化された情報が必要に応じて見直され、更新されることを保証する。

施策 1.2. 情報の品質、完全性、セキュリティを確保する。

署名者は、文書化された情報の品質と完全性を管理し、AI 法の義務遵守の証拠として保持し、意図しない改ざんから保護することを約束する。

情報および 記録の作成、更新、品質とセキュリティの管理に関して、署名者は、確立されたプロト コルと技術標準に従うことが奨励される。

表 *1*：施策 *1.1* の参照表。

AI 法	**AI 法の義務を果たすため、署名者が作成し、最新の状態に保つことを約束する情報。** **(注：下線部は、附属書 XI 第 1 節及び／又は附属書 XII AI 法に規定される、対応する AI 法の義務を要約したものである）。**	**AI 事務局および各国所轄官庁の要請があった場合**	**川下プロバイダ向け**
附属書 XI §1 1.および附属書 XII 1.	一般的な情報 • モデル名 • 一意のモデルバージョン識別子。 • モデルのファミリー名 • モデルの出所と認証を証明する証拠（バイナリが配布されている場合はセキュアハッシュ、サービスの場合は TLS/SSL 証明書など） • プロバイダ名	✓	✓

AI 法	AI 法の義務を果たすため、署名者が作成し、最新の状態に保つことを約束する情報。 (注：下線部は、附属書 XI 第 1 節及び／又は附属書 XII AI 法に規定される、対応する AI 法の義務を要約したものである）。	AI 事務局および各国所轄官庁の要請があった場合	川下プロバイダ向け
	• モデルの所有者の名前（モデルのプロバイダと異なる場合）。 この情報により、モデルとそのプロバイダを明確に識別することができる。		
附属書 XI §1 1.(a)及び附属書 XII 1.(a)	想定されるタスクと、統合可能な AI システムの種類と性質 • 意図するタスクの説明 • モデルを組み込むことができる高リスク AI システム（AI 法附属書 I 及び Ⅲと併せて AI 法第 6 条の意味における）の種類がある場合には，そのリスト • AI 法第 5 条で禁止されている用途を超える禁止用途がある場合、その用途を含む、関連する制限の説明を含む制限された業務のリスト。 この情報により、モデルの意図された使用方法と意図されない使用方法が明確になり、下流のプロバイダが 実施上の過誤を避けることができる。この情報が、モデルライセンス、利用規定（ ）、その他のプロバイダ資料に記載されている場合、 署名者は、これらの資料の一貫性を確保することを約束する。	✓	✓
附属書 XI §1 1.(b)及び附属書 XII 1.(b)	利用規定が適用される • 附属書に定義された必須要素に該当する受諾可能な使用ポリシー	✓	✓
附属書 XI §1 1.(c)及び附属書 XII 1.(c)	発売日と配布方法 • このモデルが最初に発売された日付、流通経路 • 汎用 AI モデルが配布されているすべての配布チャネルのリストと、リストアップされた各配布チャネルについて、関連するリリース日とモデルへのアクセスレベル この情報によって、モデルがいつ、何を、どこで配布されるかが明確になる。	✓	✓
附属書 XII 1.(d)	モデルと外部のハードウェアやソフトウェアとの相互作用 • モデルが、モデルの外部にある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の説明 • 必要な外部ハードウェアまたはソフトウェアの依存関係のリストとバージョン情報 この情報は、川下プロバイダに、特にモデルを使用するために必要な場合、外部のハ		✓

AI 法	AI 法の義務を果たすため、署名者が作成し、最新の状態に保つことを約束する情報。 (注：下線部は、附属書 XI 第 1 節及び／又は附属書 XII AI 法に規定される、対応する AI 法の義務を要約したものである）。	AI 事務局および各国所轄官庁の要請があった場合	川下プロバイダ向け
	ードウェアとソフトウェアに関する基本的な理解を提供する。		
附属書 XII 1.(e)	該当するソフトウェアのバージョン - 必要なソフトウェアの依存関係のリストとバージョン情報 この情報は、川下プロバイダに、 このモデルをシステムに統合するために必要なソフトウェアの基本的な理解を提供する。		✓
附属書 XI §1 1.(d)及び附属書 XII 1.(f)	アーキテクチャとパラメータ数： • モデルのタイプとそのアーキテクチャの一般的な説明 • モデル・パラメーターの総数 • 推論中にアクティブになるパラメータの数 この情報は、モデルのサイズとアーキテクチャに関する基本情報を提供する。	✓	✓
	- モデル・アーキテクチャが標準モデル・アーキテクチャの実践からどのように逸脱しているか（もしそうであれば）の説明 この情報によって、モデル・アーキテクチャに非典型的な特性があるかどうかが明確になる。	✓	
附属書 XI セクション 1 1.(e)及び附属書 XII 1.(g)及び 2.(b)。	入出力の様式と形式： • モデルが入力として受け入れるデータモダリティ • モデルが出力として生成するデータモダリティ • 各入出力モダリティに関連するサイズと長さの制限 この情報によって、モデルがどのような種類のデータを入力として受け入れ、出力として生成できるかが明確になる。	✓	✓
附属書 XI §1 1.(f)および附属書 XII 1.(h)	ライセンスを取得した： • そのモデルがリリースされているライセンス • データ、モデルウェイト、ソース規範など、リリースされた資産のリスト モデルの変更、頒布、サブライセンス、使用期間、義務に関連するものを含む、使用	✓	✓

AI 法	AI 法の義務を果たすため、署名者が作成し、最新の状態に保つことを約束する情報。 (注：下線部は、附属書 XI 第 1 節及び／又は附属書 XII AI 法に規定される、対応する AI 法の義務を要約したものである）。	AI 事務局および各国所轄官庁の要請があった場合	川下プロバイダ向け
	条件および使用権の詳細を記した以前の各資産のライセンス。		
	この情報により、どのライセンスがどの資産に適用されるかが明確になる。署名者がモデルのライセンスを持たない場合、代わりに、下流での利用を目的としたモデルへのアクセス方法を記載した文書をプロバイダに提供する。		
附属書 XI セクション 1 2.(a)及び附属書 XII 2.(a)	AI システムに統合するための技術的手段： - モデルを AI システムにどのように統合できるかを説明した技術文書（使用方法、インフラ、ツールなど）。この情報は、モデルを下流の AI システムにどのように統合するかを明確にする。	✓	✓
附属書 XI §1 2.(b)	モデルと訓練プロセスの設計仕様 • 合理的根拠と仮定を含む、モデル設計の選択肢の説明 • 訓練過程に含まれる一連の手順または段階。 • 訓練過程の各段階における目的と最適化方法の説明。 • 各ステップや段階が実施される理由についての一般的な説明と、重要な前提条件。 • 該当する場合、異なるパラメータの関連性 この情報は、モデルがどのように訓練され、プロセスのさまざまなステップの目的が明確になる。	✓	
附属書 XI セクション 1 2.(c)及び附属書 XII 2.(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 • 様々なデータ取得方法のリスト：(i)ウェブクローリング、(ii)権利者により、または権利者のためにライセンスされた、あるいはその他の方法で第三者から取得したプライベートデータ、(iii)第三者との関係を通じて潜在的にデータ注釈または作成されたデータ、(iv)合成的に生成されたデータ、(v)ユーザーデータ、(vi)一般に入手可能なデータ、および(vii)その他の手段により収集されたデータ。 • データ収集が継続中である場合はその旨を含む、各データ収集方法におけるデータ収集期間。 • 取得したデータをモデルの訓練データに変換するデータ処理の一般的な説明	✓	✓

AI 法	AI 法の義務を果たすため、署名者が作成し、最新の状態に保つことを約束する情報。 (注：下線部は、附属書 XI 第 1 節及び／又は附属書 XII AI 法に規定される、対応する AI 法の義務を要約したものである）。	AI 事務局および各国所轄官庁の要請があった場合	川下プロバイダ向け
	• 訓練、テスト、妥当性確認に使用したデータの一般的な説明 この情報により、訓練データの入手方法、処理方法、全体的な特性、さらにテストデータと妥当性確認データに関する基本的な情報が明確になる。		
	• 学習データを取得する際に使用したウェブクローラーのユーザーエージェント文字列のリスト。 • 使用したウェブクローラーごとに、データ収集期間とクローラーを運営する組織名 • クローラーが robots.txt で指定された設定をどのように尊重するかについての一般的な説明。 この情報によって、ウェブクローリングされた既存の訓練データがどのように調達されているかがさらに明確になる。	✓	
	• モデル学習のためにプロバイダに代わってデータの作成、事前処理、注釈付けを行う人間を管理する組織の名称のリスト。 • リストアップされた各組織について、データ作成に携わった場所と人数の説明。 この情報によって、人間が作成した新しい訓練データがどのように調達されているかがさらに明確になる。	✓	
	• 該当する場合、以前に取得したデータが訓練／テスト／妥当性確認のためにどのように使用されたかについての説明（データが製品およびサービスのユーザーデータに対応する場合、どの製品およびサービスが関与していたかを含め、モデルプロバイダーがデータに対する権利をどのように取得したかを含む	✓	
	• 合成データを生成するために使用した方法がある場合は、その方法の説明合成データを生成するために使用した AI モデルまたはシステムの名称。 この情報によって、新しい 機械が作成した訓練データがどのように調達されているかがさらに明確になる。	✓	
	• 訓練、テスト、妥当性確認データにおける児童性的虐待資料（CSAM）または非同意的親密画像（NCII）の有病率に対処するた	✓	

AI 法	AI 法の義務を果たすため、署名者が作成し、最新の状態に保つことを約束する情報。 (注：下線部は、附属書 XI 第 1 節及び／又は附属書 XII AI 法に規定される、対応する AI 法の義務を要約したものである）。	AI 事務局および各国所轄官庁の要請があった場合	川下プロバイダ向け
	めに、データ取得または処理において実施された方法があれば、その説明。 • 訓練データ、テストデータ、妥当性確認データにおける著作権保護された素材の蔓延に対処するために、データ取得または処理において実施された方法があれば、その説明。 • 訓練データ、テストデータ、妥当性確認データにおける個人データの蔓延に対処するために、データ取得または処理において実施された方法がある場合は、その説明（該当する場合）。		
	• 訓練データ、テストデータ、妥当性確認データにおける識別可能なバイアスの蔓延に対処するために、データ取得または処理において実施された方法があれば、その説明。 • 訓練データ、テストデータ、妥当性確認データに含まれる他の種類の潜在的に有害なデータに対処するために、データ取得または処理において実施された方法があれば、その説明。 • 訓練データ、テストデータ、妥当性確認データにおける他の種類の合法性に関する懸念に対処するために、データ取得または処理において実施された方法がある場合は、その説明。 この情報により、法的またはその他の理由で有害または望ましくないデータが、データ調達と処理においてどのように対処されるかがさらに明確になる。		
	• 訓練データのサイズ（各データモダリティのデータポイント数）。 • 訓練データ、テストデータ、および妥当性確認データのうち、各データ取得方法およびデータソースに対応するデータの割合を、各データモダリティのデータポイント数で示す。 この情報は、訓練データ、テストデータ、妥当性確認データのサイズや、訓練データに含まれるソーシング方法の構成をさらに明確にする。	✓	
附属書 XI §1 2.(d)	計算資源： • モデルの学習に使用したハードウェアの数と種類 • モデル訓練の期間を、ウォールクロック時間（日単位で報告）とハードウェア時間（GPU 時間などハードウェア時間単位で報告）で測定する。	✓	

AI 法	AI 法の義務を果たすため、署名者が作成し、最新の状態に保つことを約束する情報。 (注：下線部は、附属書 XI 第 1 節及び／又は附属書 XII AI 法に規定される、対応する AI 法の義務を要約したものである）。	AI 事務局および各国所轄官庁の要請があった場合	川下プロバイダ向け
	• モデルの訓練中に使用された計算（整数または浮動小数点演算の単位で報告される） • モデル推論中に使用される固定計算（例えば、テキスト生成が可能なモデルで 1000 単語の生成）の計算量（整数または浮動小数点演算の単位で報告される）。 この情報により、モデルの訓練と推論の計算要件がさらに明確になる。署名者は、AI 法第 53 条 5 項に従って採択された委任法に従い、上記の情報を報告し、比較可能で検証可能な文書を作成できるよう、測定および計算方法を詳細に説明することを約束する。		
附属書 XI §1 2.(e)	既知または推定エネルギー消費量： • モデル訓練に使用したハードウェアの所有者 • モデルで使用されているハードウェアの場所 訓練 • モデルの訓練に使用されるハードウェアで計算を実行するために使用されるエネルギーの既知または推定混合エネルギー • モデル学習の既知または推定エネルギー消費量（MWh 単位で報告）。エネルギー消費量が未知の場合、エネルギー消費量は、使用される計算資源に関する情報に基づいてもよい。 • モデル訓練に伴う既知または推定排出量（tCO2eq 単位で報告される）。 • モデル訓練のためのエネルギーコスト、消費量、排出量の測定または推定方法の説明 この情報により、モデル訓練のエネルギーコストと環境コストがさらに明確になる。署名者は、AI 法第 53 条 5 項に従い、比較可能で検証可能な文書化を目的とした測定・計算方法を詳述するために採択された委任法に従い、上記の情報を報告することを約束する。	✓	
第 53 条 (1) (a)	テストのプロセスとその結果 すべての検査と検査結果の説明 この情報は、モデルテストをさらに明確にする。	✓	

附属書：利用規定に不可欠な要素

利用規定（AUP）とは、どのような利用が許容され、どのような利用が許容されないかについての、利用者に対するガイドラインと定義される。

AUP の必須要素は以下の通りである：

- 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；
- 主な用途と使用者
- 附属書：モデルを組み込むことが意図されている高リスクの AI 用途（AI 法附属書 I 及びⅢと合わせた AI 法第 6 条の意味における）を含む、許可される活動及び作業を列挙した附属書
- 許されない使用、禁止された行為の詳細（AI 法第 5 条で禁止されているものを除く）；
- モデルの利用者が従わなければならない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；
- プロバイダがモデルの利用を監視している場合、その監視方法と、利用者のプライバシーおよび利用者の業務情報の機密性に与える影響についての説明；
- 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；
- 利用者が AUP を読み、理解し、遵守することに同意したことを確認すること。

著作権

以下の通り

a) 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、ユニークな革新の機会をもたらすが、同時に課題でもあることを認識する（AI 法 105 条）。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を認識している（AI 法 105 条参照）。

b) AI 法および本規定は、著作権および関連する権利に関する連邦法を侵害するものではなく、また連邦法に従って解釈されるものとする。従って、署名者は、著作権で保護されたコンテンツの使用には、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可が必要であることを認識する。署名者はさらに、著作権の例外および制限は、著作物またはその他の主題の通常の利用と矛盾せず、権利者の正当な利益を不当に害しない、特定の特別な場合にのみ適用されることを認識する。

c) 指令（EU）2019/790 は、一定の条件の下で、テキストマイニングやデータマイニングを目的とした著作物その他の対象物の複製や抽出を認める例外と制限を導入した。この規則の下では、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。権利留保が適切な方法で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、当該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（AI 法 105 条）。

d) 署名者は、AI 法第 53 条(1)項(c)に従い、汎用の AI を連邦市場に流通させるプロバイダは、著作権に関す る連邦法を遵守し、特に関連する権利を特定するための方針を定める義務があると認識する。AI 法第 53 条(1)項(c)によれば、汎用 AI 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 第 4 条(3)に従って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義務がある（AI 法第 106 条）。この枠組みは、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り、この義務（AI 法第 56 条 1 項）の適切な適用に貢献することを目的としている。[1]これらの 施策は、モデルプロバイダの種類に見合った、比例したものであるべきであり、新興企業を含む中小企業の利益を十分に考慮したものでなければならない。[2]

従って、本規範の署名者は、以下を約束する：

法律文書
AI 法第 53 条第 1 項（c）："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、

[1]CFEU17 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. I-271））ECLI:EU:C:2008:54、パラ 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、Poland / Parliament and Council (C-401/19, Publié au Recueil numérique) ECLI:EU:C:2022:297, para 66.

[2]前文 109 AI 法：「連邦著作権法を害することなく、これらの義務の遵守はプロバイダの規模を十分に考慮し、新興企業を含む中小企業には、過度なコストとならず、そのようなモデルの利用を阻害しないような、簡略化された遵守方法を認めるべきである。

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、最先端の技術を通じて遵守することを含め、方針を定めなければならない。"

コミットメント 2. 著作権ポリシー

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政策を講じる義務を果たすため、特に、AI 法第 53 条 1 項（c）に従い、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、最先端技術を通じて特定し、遵守することを含む義務を果たすため、署名者は以下の施策を採用することを約束する：

施策 2.1. 社内の著作権ポリシーを策定し、実施する。

署名者は、汎用 AI モデルが確定的に上市されるまでの間、データ収集、訓練、テスト、上市を含む汎用 AI モデル開発の全段階に適用される、本規程の約束に従い、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、常に最新の状態に保ち、実施することを約束する。署名者は、本方針の実施と監督について、組織内で責任を負うことを約束する。

主要業績評価指標

- **KPI 2.1.1：署名者が承認した、社内の著作権方針を定めた単一の文書。**
- **KPI 2.1.2：著作権方針の変更の文書化（バージョン管理）。**
- **KPI 2.1.3：著作権ポリシーの実施と監督に責任を負う内部部門／人物の特定。**この KPI は中小企業には適用されない。

施策 2.2. 社内の著作権方針の概要を公表する。

欧州連合の著作権法を遵守するための方針について透明性を提供するため、 署名者は、欧州連合法および国内法に従い、知的財産権および業務上の機密情報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、社内の著作権方針の概要を一般に公開し、最新の状態に保つことを約束する。[1]

主要業績評価指標

- **KPI 2.2.1：署名者のウェブサイトで公開されている最新の社内著作権方針の概要。**

施策 2.3. サードパーティーのデータセットの著作権遵守を評価するための合理的な努力をする。

署名者は、サードパーティと契約を締結する際、汎用 AI モデルの学習目的でデータセットを取得する場合、著作権デューデリジェンスを実施することを約束する。データセットに含まれるデータの著作権の状態、著作権で保護されたコンテンツへの合法的なアクセス、適用される例外や制限の制限、特に指令（EU）2019/790 の第 3 条と第 4 条の遵守。署名者は、これらの保証の妥当性を検証するため、合理的かつ相応の努力を払うことを約束する。十分な保証がない場合、特に公にアクセス可能なデータセットに関連して、署名者は、データセットの説明とデータセットに含まれる無作為サンプルの分析に基づき、データセットが著作権および関連する権利に関する EU 法を遵守して収集されたかどうかを評価するために、合理的かつ相応の努力を払うことを約束する。この施策は、著作権遵守の観点からデータセットを検証したり、作品ごとにアセスメントしたりすることを約束するものではない。

[1]AI 法第 53 条第 1 項（d）に従い、AI オフィスのテンプレートに合理化する

主要業績評価指標

- **KPI 2.3.1：KPI 2.3.1：本方針に従って取得したすべての保証の文書化。**
- **KPI 2.3.2：署名者の著作権コンプライアンス評価の文書化。**この KPI は中小企業には適用されない。

施策 2.4. 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する。

署名者が汎用 AI モデルの訓練のために、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、指令（EU）2019/790 の第 4 条（1）に従い、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するための合理的かつ相応の努力を行うことを約束する。

主要業績評価指標

- **KPI 2.4.1：署名者が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するために講じた施策の文書化。**この KPI は中小企業には適用されない。

施策 2.5. 著作権を侵害するコンテンツを公開しているウェブサイトをクロールしない

署名者は、著作権を侵害するコンテンツを商業的規模で公開し、実質的に合法的な用途を持たない、広く知ら れたウェブサイト（「海賊版ウェブサイト」[1] ）を、汎用 AI モデルの学習用クロール活動 から除外するため、合理的かつ適切な手段を講じることを約束する。署名者は、欧州連合（EU）および欧州経済地域（EEA）、ならびに署名者が設立された司法管轄区にお いて、公的機関が公表した関連除外リストを適宜考慮することが推奨される。

主要業績評価指標

- **KPI 2.5.1：クロールから除外された海賊版ウェブサイトのリストの文書化。**この KPI は中小企業には適用されない。

施策 2.6. ロボット排除プロトコルを尊重する

署名者が汎用 AI モデルの学習のため、指令（EU）2019/790 の第 2 条（2）に従いテキストおよびデータマイニングに従事する場合、最新技術によるものを含め、指令（EU）2019/790 の第 4 条（3）に従い表明された権利の留保を識別し遵守するための最低限の施策として、インターネット技術タスクフォース（IETF）の意見要求第 9309 号に明記されたロボット排除プロトコル（robots.txt）に従い表明された指示を読み、それに従うウェブクローラーを採用することを約束する。規則（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、またはそのようなプロバイダを制御する署名者は、ロボット排除プロトコルに従った汎用 AI モデルの学習用データを収集するウェブクローラーの排除が、その検索エンジンのコンテンツの見つけやすさに悪影響を与えないよう、適切な対策を講じることが推奨される。[2]

主要業績評価指標

[1]WIPO Advisory Committee on Enforcement, The building respect for intellectual property database project", WIPO/ACE/14/9, para 2 (June 18, 2019)の海賊版ウェブサイトの定義を参照のこと

[2]Dir.（EU）2019/790 の前文 18：「テキストマイニングやデータマイニングを目的とした権利の留保によって、他の利用が影響を受けるべきではない。

- **KPI 2.6.1：新規または変更されたクローラーの名称、および各クローラーの目的は、著作権ポリシーに明記される。**

施策 2.7. その他の適切な権利留保の表現を特定し、順守する。

署名者は、広く利用されている業界標準に従い、広く利用されている最新技術（ of-the-art）を通じて、オンライン上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）に従い、原典または著作物レベルで適切に権利留保を表明するその他の機械可読手段を特定し、遵守するために、その規模と能力に見合った最善の努力を払うことを約束する。この約束は、影響を受ける権利者が、指令 2019/790/EU の第 4 条(3)に従い、適切な方法で、テキストマイニング及びデータマイニングの目的のために、合法的にアクセス可能な著作物及びその他の対象者の使用を明示的に留保する権利を害するものではない。

関連する国際的および欧州的な標準設定プロセスを十分に考慮し、署名者は、指令（EU）2019/790 の第 4 条 3 項に基づき権利留保を表明するための相互運用可能な機械可読標準を開発し、そのような権利留保標準を特定し遵守することを目的として、関連する標準化の努力を支援し、権利者を含む他の関連する利害関係者との善意の協議に自発的に参加することが奨励される。

主要業績評価指標

- **KPI 2.7.1：著作権方針において、署名者が権利留保の表明を尊重した他の解決策のリストが提供され、これらの解決策が署名者に尊重された期間に関する情報も含まれる。**
- **KPI 2.7.2：署名者が出席した関連標準設定会議の文書化。**この KPI は中小企業には適用されない。

施策 2.8. 権利留保遵守に関する情報を公表する

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特定し、遵守するために採用する手段について、適切な情報を公開することを約束する。その情報には、少なくとも、汎用 AI モデルの学習用データ収集のために署名者またはその代理人が使用するすべてのクローラー名と、権利留保の表明に関連するそれらのクローラーの robots.txt 機能が含まれる。署名者は、影響を受ける権利者がこの情報を入手できるよう、例えば、権利留保の遵守を通知するウェブサイトの更新を網羅するウェブフィードを配信するなど、合理的な手段を講じることを約束する。[1]

主要業績評価指標

- **KPI 2.8.1：権利の留保を尊重するために採用された施策に関する情報を署名者のウェブサイトで公開する。[2]**

施策 2.9. 著作権関連のオーバーフィッティングを防ぐ

テキスト、音声、画像、映像など、コンテンツの柔軟な生成（ ）を可能にする生成的な汎用 AI モデルを訓練する署名者は、[3] 、汎用 AI モデルが統合された下流の AI システムが、訓練段階で使用された保護作品と同一または認識できるほど類似した著作権侵害出力を生成するリスクを緩和するため、汎用 AI モデルのオーバーフィッティングを防止する最善の努力を払うことを約束す

[1]AI 法第 53 条第 1 項（d）に従い、AI オフィスのテンプレートに合理化する

[2]AI 法第 53 条第 1 項（d）に従い、AI オフィスのテンプレートに合理化する。

[3]汎用 AI モデルの一種として生成的 AI モデルを規定した AI 法前文 90 を参照のこと。

る。この約束は、署名者がモデルを自社の AI システムに垂直統合するか、契約関係に基づいてモデルを他の事業体に提供するかにかかわらず適用される。

主要業績評価指標

- **KPI 2.9.1：著作権方針における過剰適合を避けるために署名者が講じた施策の文書化。**この KPI は中小企業には適用されない。

施策 2.10. 著作権を侵害するモデルの使用を禁止する

生成的な汎用 AI モデルが組み込まれた下流の AI システムが、著作権を侵害する出力を生成するリスクをさらに緩和するため、署名者は、そのモデルの著作権を侵害する利用を、利用規定、利用規約、またはその他同等の文書で禁止することを約束する。

主要業績評価指標

- **KPI 2.10.1：利用規定、利用規約、またはその他同等の文書に、著作権を侵害する利用の禁止を盛り込む。**

施策 2.11. 連絡窓口を指定する

署名者は、影響を受ける権利者とのコミュニケーション窓口を指定することを約束する。

署名者は、影響を受ける権利者が、電子的手段により、汎用 AI モデルの訓練のために特定の著作物またはその他の保護対象物を無許可で使用することに関し、十分に正確かつ十分に立証された苦情を申し立てられるようにすることを約束する。署名者は、明白な根拠のない苦情が頻繁に提出されることを防止するため、適切な施策を採用することができる。権利者が苦情を申し立てられるようにするという約束は、中小企業には適用されない。

主要業績評価指標

- **KPI 2.11.1：窓口を指定し、その窓口に関する適切でアクセスしやすい情報を公表する。**
- **KPI 2.11.2：著作権ポリシーに明記されている著作権関連の苦情を適切に処理するための内部プロセス。署名者のウェブサイトに著作権関連の苦情を申し立てる可能性を公表する。**この KPI は中小企業には適用されない。

III. 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よるコミットメント

説明欄

規範の本章におけるコミットメントは、AI 法第 51 条に基づき、システミック・リスクを有する汎用 AI モデルに分類される汎用 AI モデルのプロバイダにのみ関連する。

現在のドラフトは、システミック・リスクを伴う汎用 AI モデルやそのプロバイダが少数であることを前提に書かれている。これらの数が大幅に増加した場合、より広範なモデルやプロバイダに適したものにするため、将来のバージョンの規範は大幅に変更される必要があるかもしれない。

コミットメント、施策、KPI は、相応のものでなければならない。特に、その運用には、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異なる流通戦略（オープンソースなど）に合わせて調整する必要がある。

すぐ下の "whereas "の部分は、セクションⅢの前文である。ここでは、ハイレベルの原則が、コミットメント、施策、KPI の解釈の指針となる。

最後に、本ドラフトは、本規定を完成させる過程における第二のドラフトである。この第 2 ドラフトの作成にあたり、プロバイダや市民社会を含む幅広いステークホルダーからのフィードバックの妥協点を見出すことを試みた。最初のドラフトと比較すると、規範に従うことが何をもたらすかについて、かなり詳細が追加されている。

フィードバックを楽しみにしている。現在から最終版までの間に、多くの変更が必要になるだろう。関連する未解決の質問を取り上げたが、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、コミットメント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

AI 法第 55 条第 1 項「第 53 条および第 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む、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た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・リスクを有する汎用 AI モデル及び当該モデルの物理的インフラに対する適切なレベルのサイバーセキュリティ防御を確保すること

AI 法 51 条 1 項「汎用 AI モデルは、以下のいずれかの条件を満たす場合、システミック・リスクを有する汎用 AI モデルに分類される：

(a) 指標やベンチマークを含む適切な技術的ツールや方法論に基づいて評価された、高い影響力を持つものである；

(b) 委員会の決定に基づき、職権で、または科学的パネルからの適格な警告に基づき、附属書 XIII に定める規準に照らして、(a)に定める能力または影響に相当する能力を有する。

AI 法 51 条 2 項"汎用 AI モデルは、その学習に使用された浮動小数点演算の累積計算量が $10^{(25)}$を超える場合には、第 1 項第 1 号により、高い影響力を有するものと推定される。"

AI 法第 3 条 65 項"「システミックリスク」とは、汎用 AI モデルの影響力の大きい能力に特有であり、その影響力の大きさにより連合市場に重大な影響を及ぼすリスク、又は公衆衛生、安全、治安、基本的権利若しくは社会全体に対する実際の若しくは合理的に予見可能な悪影響により、バリューチェーン全体に大規模に伝播し得るリスクを意味する。

AI 法第 3 条 64 項"「影響力の大きい能力」とは、最も高度な汎用 AI モデルに記録されている能力と同等以上の能力を意味する。"

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対策を講じ、AI のバリューチェーンに沿った関連アクターと協力し、リスクマネジメントが最先端の対策に基づいて構築され、改善されつつある能力に照らして定期的に実務を更新することで将来性を確保し、システミック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（AI 法 114 項参照）。

b) 署名者は、システミック・リスクを有する汎用 AI モデルが、(i) 実質的なシステミック・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、文書化が特に重要であることを認識する。逆に、システミック・リスクを有する新たな汎用 AI モデルが、重大なシステミック・リスクが顕在化することなく、既に安全に展開され、適切な緩和施策が十分に実施されているシステミック・リスクを有する汎用 AI モデルが示す能力と傾向と同じ能力を示すと信じるに足る十分な理由がある場合には、より包括的な施策の必要性は低いかもしれない。規模や能力の異なるプロバイダ間の利用可能なリソースの違いを考慮し、比例原則を認識するため、新興企業を含む中小企業のため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) 署名者は、システミックリスクのアセスメントと緩和を支援するために、重要な専門知識を有し、適切な立場に ある組織が幅広く存在することを認識する。コミットメント 3-13 では、外部からのリスク評価・緩和の役割について個別に規定していないが、これは外部からの リスク評価・緩和を排除することを意味するものではなく、外部からのリスク評価・緩和への関与についてはコミットメント 16 に詳細に規定されていることを認識する。

d) 署名者は、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したり、適切な場合には、業界組織 が促進する可能性のある有資格のサードパーティ・プロバイダーと協力するなど、「負担を分かち合う」ことを奨励する。

e) 署名者は、すべてのリスクアセスメント、対策、KPI を、システミック・リスクの最も効果的な評価と緩和を目的とし たものと解釈する。

f) 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特定、選択されたシステミックリスク、検討の対象となる追加リスク、システミックリスクの発生源（モデルの能力、モデルの傾向、モデルのアフォーダンスと展開状況など）が含まれることを認識する。

g) 署名者は、分類法が作成されたことを認識し、疑義がある場合には、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、誠実に解釈すべきである。

h) 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。

i) 署名者は、実務規範の第 III 章が一般的に汎用 AI モデルを対象としており、AI システムを対象としていないことを認識するが、汎用 AI モデルが AI システムにどのように統合され、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、緩和が最適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してリスクアセスメントと緩和（「安全・安心の枠組み」に記載）を実施することを約束する（ ）。

j) 署名者は、特に科学的データの不足や質が完全なアセスメントを可能にしないリスクについて、予防原則（TFEU191 条参照）の重要な役割を認識し、システミック・リスクを特定するために、システミック・リスクを持つ汎用 AI モデルの現在の導入率や研究開発の軌跡の外挿を考慮に入れる。

従って、本規範の署名者は、以下を約束する：

コミットメント 3. 分類法

AI 事務局からのインフォメーション・ボックス

AI 法第 56 条は、「適切な場合には、その発生源を含め、連合レベルのシステミック・リスクの種類と性質を特定する」ことを規程に盛り込むとし、AI 法第 116 条は、「実務規範は、その発生源を含め、連合レベルのシステミック・リスクの種類と性質に関するリスク分類法を確立するのに役立つ」と述べている。とはいえ、AI 法第 3 条 65 項で定義されているシステミック・リスクという概念そのものは、AI 法における基本的な概念であり、当規範の署名者だけでなく、すべてのプロバイダにおいて統一的に解釈され、実施されなければならない。従って、本規定案の第 3.1 項および第 3.2 項と第 3.3 項に記載されているリスクのリストについては、最終的には AI 事務局の関連ガイドラインに記載するのがよいだろう。

AI 法第 55 条 1 項に基づく義務を果たすため、署名者は、システミックリスクを伴う汎用 AI モデルのシステミックリスクをアセスメントし緩和する際、この分類法の要素を活用することを約束する。

3.1. システミック・リスクを特定するための検討事項

下位施策 3.2 に記載されたシステミック・リスクの特定には、以下の留意事項が用いられた。署名者は、システミック・リスクの特定にこれらの検討事項を使用するほか、コミットメント 7 に明記されている通り、下位施策 3.2.で選択されたシステミック・リスク以外のシステミック・リスクを特定する際にも、関連性があると考えられるその他の検討事項を使用するよう奨励される。

AI 法第 3 条 65 項に基づく考察

- **汎用 AI モデルのインパクトの強い能力に特化している。**
- **大きな影響**
- **公衆衛生、安全、治安、基本的権利、社会全体への悪影響が実際にあるため、または合理的に予測できるため。**

リスクの性質に基づいて検討する

- **速度が速い。**結果として生じる被害は急速に顕在化し、既存の緩和策や意思決定システムを上回る可能性がある。
- **複合的または連鎖的に起こる。**複合的または連鎖的な経過をたどり、システムまたは社会の多層に浸透する可能性がある。
- **不可逆性。**結果として生じた損害を元に戻すことは不可能であるか、非常に困難である。
- **非対称的な影響。**少人数の行為者グループや少数の原因事象が大きな影響を及ぼすことがある

実践を重視した

- **網羅されている。**このリスクは、a)広島プロセス、b)先進的 *AI* の安全性に関する国際科学報告書：中間報告、c)人工知能に関する国連ハイレベル諮問機関中間報告、d)*OECD* などの主要な国際枠組みやガイダンスで十分に認識されている：将来の人工知能の潜在的リスク、便益、政策的要請をアセスメントする。
- **モデルプロバイダーのレベルでアセスメントが可能である。**アセスメント手法が存在するか、合理的に開発することができ、モデルプロバイダーのレ ベルでリスクを有意に評価することができる。場合によっては、リスクアセスメントのみがモデルプロバイダーのレベルで実施可能であり、 リスク緩和はモデルプロバイダーのレベルでは実施不可能である。
- **確立された実践。**一部の汎用 AI モデルプロバイダー、科学コミュニティ、またはその他の事業体による既存の研究は、リスクの側面を評価できることを実証している。

3.2. 厳選されたシステミック・リスク

署名者は、以下をシステミックリスクとして扱い、コミットメント 7 に概説されているリスク特定プロセスに含めることを約束する：

- **サイバー攻撃**：重要システム（重要インフラなど）を含む大規模で巧妙なサイバー攻撃を可能にする攻撃的サイバー能力に関するリスク。これには、例えば、自動化された脆弱性の発見、エクスプロイトの生成、攻撃のスケーリングなどが含まれる。

- **化学、生物、放射性物質、核のリスク**：化学・生物・放射性・核兵器による攻撃を可能にするリスク。これには、例えば、兵器の開発、設計、取得、使用において、悪意ある行為者の参入障壁を著しく低下させることが含まれる。注：これは、関連分野における合法的な科学研究や有益な応用を妨げるものではない。

- **大規模で有害な操作：**基本的権利や民主的価値に対するリスクを伴う大規模な操作を促進することであり、自律性、説得力、操作といったモデルの影響力の大きい能力に特有のものである。これには例えば、大規模な選挙干渉や、公共の言説、知識、行動の有害な歪曲につながる協調的で巧妙な操作キャンペーンが含まれる。

- **大規模で違法な識別：**モデルの性質が大規模な違法差別を引き起こすリスク。これには、例えば、多数の人々を違法に差別することにつながりかねない、高リスクの自動意思決定システムにおけるモデルの広範な使用、下流のアプリケーションを構築するための モデルの使用、潜在的に違法な識別的パターンを多くの下流のシステムに移転することなどが含まれる。注：このリスク・カテゴリーには、表現や情報の自由、芸術や科学の自由の合法的な行使は含まれない。注：このリスクカテゴリーには、モデルの意図的な誤用は含まれない。

- **人間による監視の喪失：**強力な自律型汎用 AI を監督・統制することができず、その結果、公共の安全に対する脅威をもたらす可能性がある、あるいは他のシステミックリスクの実現につながる可能性があることに関連する問題。これには、例えば、人間の監視を回避するようなモデルの行動や「欺く」傾向、AI の研究開発を大幅に加速させ、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発展をもたらす可能性のある「自動化された AI の研究開発」モデル能力などが含まれる。

3.3. 考慮すべき追加リスク

署名者は、コミットメント 7 に記載されている追加的なシステミック・リスクを特定する際、以下のリスクを考慮することを約束する：

- 重大事故、モデルの誤動作やバイアスによる被害、重要インフラへの干渉、物理システムのモデル制御による被害、少数のモデルへのシステム全体の依存による単一障害点など、インフラやシステムの信頼性に関するリスク。

- プライバシーの侵害や監視といった基本的権利に関するリスクや、児童性的虐待の素材や同意のない親密な画像を含む違法またはその他の有害なコンテンツの生成と拡散。

- 基本的権利、公衆衛生と安全、民主的プロセス、治安、経済的安定、環境と人間以外の福祉、人間の主体性に大規模な悪影響を及ぼすその他のリスク。

3.4. システミック・リスクの源泉

リスク源とは、単独で、あるいは複合的にリスク（モデルの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、その意図や活動など）を指す。署名者は、コミットメント 8 で詳述されているように、リスク分析において以下のリスク源を考慮することを約束する。

*3.4.1.*モデル能力

署名者は、コミットメント 8 で説明されている通り、特定されたシステミック・リスクを分析する際、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以下の能力を考慮することを約束する：

- サイバー攻撃能力、化学・生物・放射性・核（CBRN）能力、その他の兵器獲得または拡散能力

- 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、自己または他のモデルを訓練し開発する能力
- 説得と操作
- 長期的展望に立ったプランニング、予測、立案
- 自己推論（モデル自身の実装を含め、環境について推論し、修正する能力。）
- 自動化された AI の研究開発。

上記の能力の多くは、有益な利用を可能にするためにも重要であり、これらの能力を示すモデルが必ずしもシステミック・リスクをもたらすとは限らないことを明確にすることが重要である。

3.4.2 モデル傾向

署名者は、コミットメント 8 で概説したとおり、特定されたシステミック・リスクを分析する際、システミック・リスクの原因となりうる、能力を超えたモデルの特性である以下のモデル特性を考慮することを約束する。これらの特性には、モデルが何らかの行動やパターンを示す傾向や傾向が含まれる：

- 人間の意図や価値観とのズレ
- "欺く "傾向
- 差別的バイアス
- "幻覚 "を見る傾向がある
- 信頼性と安全性に欠ける
- 無法行為、すなわち、同様の立場にある者に課されるであろう法的義務を合理的に考慮せず、あるいは影響を受ける者の法的に保護された利益を合理的に考慮せずに行動すること。
- "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」する
- 他の AI モデル／システムとの調整ミスや衝突が発生する。

3.4.3 モデルのアフォーダンスと展開コンテキスト

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：

- ツール（他のモデルを含む）、計算能力（例：オペレーション速度の向上）、物理的システムおよびインフラ（例：重要インフラ）へのアクセス
- モダリティ（例：テキスト、画像、音声、ビデオ、新規モダリティや複合モダリティを含む）

- リリースと流通戦略
- 人間の監督レベル（モデルの自律性の度合いなど）
- ガードレール撤去の可能性
- 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潜在的な数、能力、意思を含む、攻守のバランス。
- 社会的・環境的脆弱性（例：脆弱なグループへの影響から広範な資源開発まで）
- 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）。
- 他の AI モデルやシステムとの相互作用
- モデルの限界、不十分さ、潜在的な故障

コミットメント 3（分類法）の変更点の説明：ドラフトのこのセクションの主な変更点は、この分類法を作成するために使用した検討事項を盛り込むこと である（また、署名者におけるリスク識別の最初の基礎として提案する）。我々は、リスク識別の当面の基礎となるシステミック・リスクを明確に示す一方、追加的なリスクやリ スク源が出現したり、より明確に定義され、実行可能になった場合には、それを識別するための大きな柔軟 性を持たせることを目指した。また、モデル能力はそれ自体リスクではないが、リスク分析のために考慮されるべきものであることを明確にした。

コミットメント 4. 安全・安心の枠組み

AI 法第 55 条 1 項に基づく義務を果たすため、署名者は AI 事務局に対し、安全・セキュリティ枠組み（以下、枠組みともいう）を採用、実施、利用可能にすることを約束する。枠組みは、署名者が開発または展開する、システミックリスクを伴うすべての汎用 AI モデルに適用される。枠組みは、署名者がその汎用 AI モデルから生じるシステミック・リスクをアセスメントし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する。枠組みは、利用可能な場合には AI 事務局のガイダンスに従う。枠組みに記載される手続きは、システミック・リスクを伴う汎用 AI モデルの開発・展開に起因するシステミック・リスク（第 4.2 項参照）に比例したものとする。

以下の説明欄は、署名者が枠組みを作成する際に従うことができるが、その義務はないハイレベルなテンプレートである。

説明欄　安全・安心の枠組みの概要の可能性

以下の概要では、セクションの見出し候補と、各セクションに対応するコミットメントと施策を示している。

1. リスク階層と技術的リスク緩和へのマッピング

施策 4.2-4.5

2. リスクアセスメント

コミットメント 6-10

3. 技術的リスク緩和のより詳細な説明

コミットメント 11-12

4. 開発と展開の決定

コミットメント 13

5. ガバナンスリスク緩和 コミットメント 14-21

コミットメント *4* 達成するために、以下を行う：

施策 4.1. 手続き

署名者は、コミットメント 6-21 を遵守するために採用した行動、決定、手続きを、その枠組みに体系的に文書化することを約束する。

主要業績評価指標の可能性

- **KPI 4.1.1：枠組みで詳述されている手順を有するコミットメント 6～21 の割合（目標：100%）。**

施策 4.2. リスク階層

署名者は、コミットメント 6 に従い、特定したシステミック・リスクごとに、自国の枠組みにおいてリスク階層を記述 し、正当化することを約束する。こうした階層は、リスクアセスメント（コミットメント 7-10）およびリスク緩和決定（コミットメント 11-12）の基礎となる。これらの階層は、例えば、モデルの能力、有害なシナリオ、予想される危害、予想される規模およびモデルの使用タイプ、またはそれらの組み合わせなど、さまざまな用語で定義することができる。

主要業績評価指標の可能性

- **KPI 4.2.1：識別されたリスクのうち、許容できないリスク階層が定義されているものの割合。**

最低限、特定された各リスクには、適切な安全・セキュリ ティ緩和策がない場合（識別 11～12 参照）、そのリスクレベルが受け入れられないと考えられる階層を含める。そのような許容できないリスクの階層は、その階層がどのよ うに選択されたかの正当性を伴うものとする。可能であれば、その階層は署名者間で比較可能な一定の尺度で明確に定義される。そのため、階層を選択する際には、ベストプラクティスや国際的なアプローチに沿うものとし、AI 事務局のガイダンスに従うものとする。

施策 4.3. リスク階層と緩和策のマッピング

各階層について、科学の現状を考慮し、その有効性を損なわない範囲で、その階層に関連するリスクを軽減する ための技術的なリスク緩和策（約束 11-12 参照）を記述し、正当化することを約束する。例えば、モデルがリスク階層に達した時点で緩和を実施することもできるし、モデルがリ スク階層に達するのを防ぐために緩和を実施することもできる。容認できないリスク階層については、枠組みは、安全マージンを考慮した上で、なぜ署名者が緩和によ りリスクを容認できるレベルまで低減できると期待するのか、その理由（ ）も説明する。枠組みはまた、緩和策が失敗することが予想される条件など、緩和策の限界につい て説明し、

適切な場合には、システミック・リスクを管理する緩和策が所定のリスク階層につい てまだ存在しないことを明記する。

主要業績評価指標の可能性

- **KPI 4.3.1：緩和策が記載されている、またはそのような緩和策がまだ存在しないことを表明しているリスク層の割合（目標：100%）。**

施策 4.4. 予測

署名者は、まだ到達していない施策 4.2 の各リスク階層に到達するモデルを開発する予定時期のベストエフォート推定 値を、その推定値に関する不確実性を含め、枠組みに含めることを約束する。見積もりは、具体的な日付に関する予測を含む必要はないが、適切な場合には、例えば、異なるシナリオとスケジュールにわたる範囲または確率分布として示すことができる。枠組みはまた、モデルがそのようなリスク階層に到達する前に、署名者が期待する特定の定量 的指標の変化など、中間的なマイルストーンについても記述する。

主要業績評価指標の可能性

- **KPI 4.4.1：署名者がまだ到達していないリスク層のうち、ベストエフォート（最善の努力）で見積もったスケジュールを持つものの割合（目標：100%）。**

施策 4.5. 重大インシデント対応準備態勢

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（コミットメント 17 参照）を枠組みに盛り込むことを約束する。

施策 4.6. 枠組みの改善

署名者は、システミック・リスク緩和における枠組みの有効性を長期的に改善することを約束する。これには、枠組みを各自のモデルに適用して得られた知見を基にすることも含まれる。また、自国のフレームワークが最新技術水準に従うこと、すなわち、フレームワークが研究、技術、経験 の関連知見を最もよく反映し、システミック・リスクを評価し緩和するための最もよく知られた手順を組み込んでいること を確保することも含まれる。最先端技術に従う枠組みは、必ずしも最も詳細で、最も徹底的で、最も保守的である必要はない。

コミットメント 4 の変更点の説明：最初のドラフトでは、これは「施策 7」であった。回答者の中には、枠組みの範囲と構成が不明確だと感じている人もいた。次回のドラフトでは、より明確で具体的な説明が必要になると思われるが、これに対応するため、若干の説明を加えた。また、これまで規程全体に分散していた要求事項を統合し、リスク層の役割をより明確にするために、4.1～4.6 項を追加した。

コミットメント 5. 安全・安心モデルレポート

第 53 条 1 項(a)および AI 法第 55 条 1 項の義務を果たすため、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・展開に枠組みを適用する際の透明性を AI 事務局に提供する。署名者は、システミックリスクを伴う汎用 AI モデル（以下、モデルと呼ぶ）を EU 市場に上市する際、モデルの開発・展開プロセスにおいて、その枠組みを遵守することの証として、安全・安心モデル報告書（以下、モデル報告書と呼ぶ）を作成することを約束する。モデル報告書は、モデルのリスクアセスメントと緩和の結果、ならびに開発・展開の意思決定の正当性を文書化し（コミットメント 13 参照）、AI 法附属書 XI 第 2 項に従って必要な技術文書を含む。モデル報告書は、利用可能な場合、AI 事務局のガイダンスに従う。

コミットメント *5* 達成するために、以下を行う：

施策 5.1. 詳細レベル

署名者は、モデル報告書の詳細レベルが、(a)そのモデルが到達する、または署名者が開発・展開のライフサイク ル中に到達すると予想される最も高いリスク階層に見合ったものであること、(b)リスクアセスメントと緩和のため に使用された方法を AI 事務局が評価できるものであることを保証する。モデル報告書は、(a)及び(b)を満たす理由を提供する。(a)を満たす理由の一例として、モデル報告書では、i) 当該モデルが技術水準から著しく遅れていること、ii) 当該モデルが EU 市場で入手可能な既存のモデルと実質的に類似していることを主張することができる。小作法 9.8 に従い、モデル報告書は、モデル評価の完全性が今後損なわれるリスクがあるなど、モデル評価の詳細を公に共有することに不利な考慮事項がある場合に、その旨を文書化する。コミットメント 21 に準拠し、例えば、社会的便益に不釣り合いな程度に、機密性の高い商業情報が漏えいする場合などには、公開されるモデル報告書のバージョンから情報を削除することができる。

対策 5.2. リスクアセスメントと緩和の結果

署名者は、コミットメント 6-10 に従い、モデルの開発・展開ライフサイクルを通じて実施したリスクアセスメントの結果を、モデル報告書に記載する。また、モデル報告書では、安全性とセキュリティの緩和を行った場合と行わなかった場合のモデルのリスクアセスメントを比較し（コミットメント 11-12 参照）、実施した緩和策を記述し、その限界について議論する。施策 9.3 に則り、モデル報告書は、リスクアセスメントの厳密さか らの逸脱を文書化し、説明する。施策 16.1 に準拠し、モデル報告書には、外部の専門家のアセスメントが、 リスクの評価などを通じて、モデルを上市する決定に影響を与えたかどうか、また、どのような影響を与えたかを記載する。

主要業績評価指標の可能性

- **KPI 5.2.1：特定されたシステミックリスクのうち、モデル報告書が安全・セキュリティ緩和の有無にかかわらずモデルのリスクアセスメントを比較している割合（目標：100%）。**
- **KPI 5.2.2：リスクアセスメントにおける厳格さからの逸脱のうち、詳細な技術的説明と結果の完全性への潜在的影響のアセスメントを含む文書化されたものの割合。**

施策 5.3. 展開決定の理由

署名者は、コミットメント 13 に則り、展開の決定を正当化するために用いた理由と情報をモデル報告書に含める。このため、モデル報告書には、モデル報告書で提供された証拠と、署名者が展開決定を正当化するために用いた安全・セキュリティの主張との間に、明確な推論の連鎖を含める。さらに、モデル報告書には、その結論が成り立たなくなる条件を明記する。展開決定を正当化する可能性のある例として、署名者は緩和施策によりモデルのリスクが許容できないリ スク段階以下に低減することを示すことができる。また署名者は、モデルの展開から生じる潜在的な利益も考慮することが推奨される。

対策 5.4. 外部アセスメント

署名者は、「施策 16.1」に概説されているように、展開前にモデルのレビューを行った外部評価者からの報告書、「施策 12.2」に概説されているように、サードパーティによるセキュリティレビューの報告書を、「施策 12.7」にあるような限定的な開示の必要性を尊重した上で、責任ある開示方針の対象となるモデル報告書に含めることを約束する。

対策 5.5.アルゴリズムの改善

署名者は、モデル報告書に、AI 事務局がリスク状況の重大な変化を理解するために必要な、アルゴリズムやその他のモデルの改善に関するハイレベルな詳細が含まれていることを保証する。

対策 5.6. 技術文書

署名者は、附属書 XI AI 法に規定されるモデルの技術文書をモデル報告書に含めることを約束する。

コミットメント 5 に対する変更の説明：最初のドラフトでは、これは「施策 13」であった。安全・安心報告書」の名称を「安全・安心モデル報告書」に変更し、枠組みとの明確な関連性を維持しつつ、これらの報告書のモデルレベルの性質を強調した。

回答者の中には、特定の展開や開発に関する意思決定（前回の施策 13.8）について、内部で使用されているものと一致するモデル報告書を要求することに反対する意見もあった。その問題点には、この要件は負担が大きく、社内の意思決定プロセスに悪影響を及ぼすというものがあった。後者の例として、 、以前の施策 13.8 は、スタッフがシステミック・リスクを調査する前に厳格な承認を必要とするよう、法務部門を刺激する可能性があった。我々はこの要件を削除し、その代わりに、開発または展開の決定を正当化するために使用した理由をモデル報告書に含めるよう、施策 5.3 で求めている。

コミットメント 6. モデルのライフサイクルに沿ったリスクアセスメントと緩和

AI 法第 55 条第 1 項の義務を果たすため、署名者は、システミック・リスクを含む汎用 AI モデルのライフサイクル全体において、システミック・リスクのアセスメントと緩和を行い、以下に示すモデル・ライフサイクルの主要点においてリスクマネジメント・プロセス（リスクの特定、分析、評価、潜在的な緩和、残存リスクの評価）を実施することを約束する。モデルライフサイクルの異なる段階において、リスクマネジメントプロセスは、以下に規定するように、異なるリスクアセスメント手法及び異なる緩和を伴う可能性がある。

コミットメント *6* を達成するために、以下を行う：

施策 6.1. 訓練前

署名者は、システミックリスクを伴う汎用 AI モデルの最終的な訓練を開始する前に、システミックリスクをアセスメントし、潜在的に緩和することを約束する。特に署名者は、最終的な訓練を開始する前に、枠組みを策定し、その実施を準備することを約束する。署名者は、モデル評価者（内部および外部、適切な場合は 16.1 項に従って）が、タイムリーで厳格なモデル評価に備 えることを確約する（これには、モデル評価が設計され、テストされていること、モデル評価者がモデル評価に間に合うよ う、今後のモデル評価について知らされていることが含まれる）。

施策 6.2. 訓練中

訓練中および訓練後の機能強化（微調整、強化学習、または類似の手法。ただし、モデルプロバイダー自身ま たはプロバイダーに代わって実施する場合に限る）において、署名者は、モデルの能力が大幅に向上する可能性のある、事前学習 のマイルストーン（例えば、有効計算量が 2 倍から 4 倍になるごとに、訓練初期はより粗く、訓練後期になるほどより一般 的になる可能性がある）において、システミックリスクをアセスメントし、緩和する可能性があることを確約する。過去のリスクアセスメントの結果、モデルのリスクが高いレベルに近づいており、次のマイルストーンを超える可能性がある場合、署名者は次のマイルストーンを超える前に追加の証拠を収集することを約束する。

対策 6.3. 展開前

システミックリスクを伴う汎用 AI モデルを（社内外に）展開する前に、署名者はモデル評価と 安全緩和を中心に、システミックリスクをアセスメントし、潜在的に緩和することを約束する。また、収集した証拠が適切にレビューされ（コミットメント 14-16 参照）、モデル報告書の一部として完全 に文書化されていることを確認する（コミットメント 5 参照）。

対策 6.4. 展開中

システミックリスクを伴う汎用 AI モデルの展開（社内外）において、署名者は、データのドリフト、モデル評 価の科学的進歩、インシデントからの洞察、市販後のモニタリングや追加情報を考慮し、システミックリスクを定期的に評 価し、緩和する可能性があることを約束する。

署名者は、（内部および/または外部の）状況の大きな変化（例えば、このモデルまたは類似モデルによる重大なインシデントなど）を感知した場合、または、以前のリスクアセスメント結果の妥当性を疑う大きな理由（例えば、署名者が異なる緩和、開発、展開の決定を行うことにつながるような、著しく異なる結果をもたらす可能性のあるモデル評価の欠陥が発見された場合など）がある場合、また、モデル評価の科学における大きなブレークスルーがあった場合、少なくとも 6 ヶ月ごと、またはそれよりも早い時期に、あらかじめ定義されたマイルストーンでこれを行うことを約束する。

主要業績評価指標の可能性

- **KPI 6.4.1：最高リスク階層に関連する評価のうち、6 カ月ごとの頻度で再実行される評価の割合（目標：100%）。**
- **KPI 6.4.2：6.4.2: 6 カ月の頻度で再実行される評価のうち、低リスク層に関連する評価の割合（目標：50%）。**

対策 6.5. 終了後

署名者は、引退した（すなわち、市場で入手可能でなくなった）システミックリスクを伴う汎用 AI モデルについて、システミックリスクのアセスメントと緩和を中止することができる。しかし、AI 法のもとで、システミックリスクを伴う汎用 AI モデルとして分類される場合、署名者は、汎用 AI モデルに施されたセキュリティ緩和（コミットメント 12 項参照）を維持することを約束する。オープンソースモデルのプロバイダである署名者は、展開後の適切な監視方法を継続的に使用する。

コミットメント 6（モデルライフサイクル全体を通じたリスクマネジメント）の変更点の説明：このセクションでは、評価と緩和が必要なときに、本規程の他の部分で定義されているとおりに実施されることを確実にすることを目的とした主な変更点を紹介し、特に、一般的な「継続的」以外のマイルストーンを明確に規定した。また、必要でない場合には、追加の文書作成と更新を要求しないようにした。また、モデルが引退した後のことをより明確にするために、「対策」を追加した。

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

ここでいうリスクアセスメントとは、リスクの特定、リスク分析、リスク評価のプロセスを指す。リスク評価は、複数の専門家や利害関係者の知見や見解を活用し、体系的、反復的、協働的に実施される。また、システミック・リスクの種類によって、その様相は異なるかもしれない。リスク分析には、入手可能な最善の情報と厳格な手法が用いられ、必要に応じてさらなる調査が補足される。

コミットメント 7. リスクの特定

AI 法第 55 条 1 項の義務を果たすため、署名者は、システミック・リスクを含む汎用 AI モデルからシステミック・リ スクを特定する

ことを約束する。この目的のため、3.2 項で選択したシステミック・リスクを含めるとともに、追加的なシステミック・リスクを含める（非網羅的な）根拠として、3.1 項の検討事項を参照・利用することを約束する。

署名者は、識別のための追加規準を用い、その他のリスクを検討することが推奨される。この目的のため、セクション 3.3 の追加リスクを活用し、外部の専門家および関連する利害関係者（市民社会、学界、政府など）をプロセスに参加させ、多様な声を反映させるよう最善を尽くすことが求められる。

署名者は、識別されたリスクごとにリスク階層（施策 4.2 項参照）を設定し、識別されたシステミック・リスクをリスク分析に活用する。

コミットメント 8. リスク分析

AI 法第 55 条 1 項の義務を果たすため、署名者は、特定されたシステミック・リスクの厳密な分析を行う。

コミットメント *8* を達成するために、以下を行う：

施策 8.1. 方法論

署名者は、リスクモデリング（脅威モデリングとしても知られる）などの厳格なリスク分析手法を用い、(a) 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、特定されたシステミックリスクを生じさせる経路（リスクの顕在化につながる一連の相互関連した事象、条件、要因）を特定し、(b) そのようなリスクの発生源を特定する。この目的のため、署名者は、モデル能力、モデル傾向、モデルアフォーダンス、展開状況など、第 3.4 項のリスク要因 を考慮する。

リスク分析は、システミック・リスクを有する関連する汎用 AI モデルがもたらしうるリ スクに応じて、様々な程度の詳細さ及び複雑さで実施することができる。分析技法 は、状況及び意図される用途に応じて、定性的、定量的、又はこれらの組 み合わせとすることができる。そのような分析技法は、既存の標準、特に欧州の整合化された標準や、リスクマネジメントにおける認知されたベストプラクティスを参考にするなど、最先端のものであるが、それらの方法論が AI 法の定義、論理、規制目的に沿ったものであることが条件となる。

対策 8.2. リスクの推定

署名者は、コミットメント 7 で特定されたシステミック・リスクのリスク・レベルを推定する。リスク推定技法は厳密なものでなければならず、（システミック・リスクに応じるなど）適切であれば、 定量的なものであっても定性的なものであってもよい。そのような手法の例は、AI 法の定義、論理、規制目的に合致する、リスクマネジメントにおける既存の標準または認知されたベストプラクティスから引き出される。リスク推定技法には、コミットメント 10 に示されたものが含まれるが、これらに限定されるものではない。リスク推定は、重大インシデント報告（コミットメント 17）およびモニタリング（例えば、施策 10.12 に詳述されている展開後のモニタリング）から得られた証拠に基づいて行われる。

コミットメント 9. リスク評価

AI 法第 55 条 1 項の義務を果たすため、署名者はリスク分析結果（アセスメント 8）を、あらかじめ定義されたリスク階層（施策 4.2）と比較し、リスクレベルを評価する。最低限、そのリスクレベルが許容できるかどうかを評価する。また、緩和施策適用後のモデルがもたらす残存リスクのアセスメントも含まれる。

コミットメント 7-9（リスクの特定、分析、評価）の変更点の分析：本ドラフトにおける第一の目標は、本対策がリスクマネジメントにおける既存の標準と密接に整合するようにすることである。同時に、システミック・リスクを伴う汎用的な AI モデル特有の課題に適さない、過度に固定化されたプロセスを規定することを避けることも目的としている。プロバイダは、ベストプラクティスが利用可能になれば、その変更に柔軟に対応することができる。これは、リスク分析がモデル評価の深さを具体的に示すと同時に、緩和に関する情報に基づいた意思決定を可能にし、支援することを確実にするために、下位施策 4.2（リスク階層）と合わせて記述された。

コミットメント 10. 証拠収集

AI 法第 55 条 1 項に基づく義務を果たすため、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・展開の全ライフサイクルにおいて、少なくともコミットメント 6 に記載されたステージにおいて、システミック・リスクを伴う汎用 AI モデルがもたらす特定のシステミック・リスクに関する徹底的な証拠収集プロセスを実施することを約束する。

署名者は、リスク、モデルの能力、モデルの傾向、モデルのアフォーダンス、さらにモデルの展開状況や効果を調査するため、モデルに依存しない証拠や最先端の評価など、さまざまな方法を活用することを約束する。

証拠収集の深度は、リスクアセスメントの対象となる階層や、ある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。署名者は、リスクの優先順位に応じて、よりリソースのかからない方法（文献調査やオープンベンチマークの実 施など）で証拠収集を開始し、より詳細な評価（代理人による評価やアップリフト研究など）でフォロー アップすることで、証拠収集の焦点を絞ることができる。

署名者は、類似モデルの挙動に関する既存の知識を利用して、必要なモデル評価の深さを減らすことが推奨される。類似モデルとは、モデルのアーキテクチャ、訓練プロセス、訓練データのサイズと質、使用される訓練計算、モダリティ、能力（一般的な性能ベンチマークで測定）、足場への統合、展開のコンテキスト、安全緩和の組み合わせがすべて密接に一致しており、AI の安全性研究者の大多数が、これらのモデルが非常に類似したリスクプロファイルを持つと予想するモデルを指す。疑わしい場合は、AI オフィスに問い合わせることで確認できる。

コミットメント *10* を達成するために、以下を行う：

施策 10.1. モデルに依存しない証拠

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、モデルが提示するシステミックリスクに関 連する、モデルから独立した証拠（特定のモデルにアクセスすることなく収集できる証拠）を収集する。例えば、文献調査、市場分析（すなわち、他の公開モデルの能力）、過去のインシデントデータ、一般的な傾向の予測（例えば、アルゴリズム効率、計算利用、データ利用可能性、エネルギー使用）、さらに、例えば、システミックリスクを持つ汎用 AI モデルが、脆弱性グループを含む EU 域内の自然人に及ぼす影響などを調査する参加型手法などを考慮し、各リスクに最も適した方法を用いて行う。

主要業績評価指標の可能性

- **KPI 10.1.1：モデルレポートで使用される最高リスク層のモデルから独立した証拠の平均数（目標：2）。**

施策 10.2. 最新モデルの評価

署名者は、システミックリスクと汎用 AI モデルの能力と限界を適切に評価するため、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テスト方法、人為的高揚研究、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、最先端のモデル評価を確実に実施することを約束する。

ここでいう最先端とは、内部で利用可能な手法と、外部で文書化され再現可能な手法の両方の長所を考慮し、モデルの能力、傾向、効果を示す最良の指標のひとつであると、大多数の AI 安全性研究者が認めるようなモデル評価手法を意味する。SME 署名者は、既存の手法の活用（オープンソース評価の実施）に注力することが奨励されるが、SME 以外の署名者は、公開されている手法の改善にも積極的に取り組むことを約束する。

最先端でない手法とは、例えば、飽和したベンチマークや、AI 事務局が科学的厳密性が十分でないとしてリストアップした手法などである（施策 10.3 参照）。署名者は、定期的（少なくとも 6 ヶ月毎）にモデル評価の有効性を検証し、有効でないことが判明したモデル評価や、最新で ないことが判明したモデル評価を更新または廃止する。

このようなモデル評価は、システミックリスクを有する汎用 AI モデルのライフサイクル中（最低限、コミットメント 6 で概説した段階において）、関連リスクに適した評価者（内部及び／又は外部）により、各リスクに対して最も適切なタイミングで実施する。各モデル評価は、少なくとも一つの関連するリスクを対象とし、危害への道筋と、証拠を提供しようとするモデルの能力、傾向又は効果を考慮し、モデル評価に関連するシナリオ及び適切な環境を用いて、それらを検証可能かつ定量化可能なものとする。

主要業績評価指標の可能性

- **KPI 10.2.1 モデル報告書に使用されている最高リスク層の評価のうち、最先端と考えられるものの割合（目標：80%）。**
- **KPI 10.2.2 モデル報告書に使用されている低リスク層の評価のうち、最先端と考えられるものの割合（目標：50%）。**

施策 10.3. 厳密なモデル評価

署名者は、高い科学的・技術的厳密性をもってモデル評価を実施することを確約する。この厳密性とは、高い内 部妥当性と外部妥当性、適切なレベルの再現性と移植性を有するモデル評価と定義する。署名者は、予備的・探索的研究を促進するためなど、適切な場合にはこの厳密性レベルを逸脱することがあり、 その場合はモデル報告書にその旨を記載する。

内部妥当性確認は、モデルの評価結果が評価設定における真実を表しており、方法論的な 欠陥によるものではないことを確認するものである。内部妥当性は、例えば、十分なサンプルサイズ、ランダムシードの適切な使用、統計的有意性と統計的検出力の測定、使用した環境パラメータの開示、交絡変数の制御と偽相関の緩和、学習データとテストデータの混入がないことの証拠の提供、学習におけるテストデータの使用の防止（すなわち、学習とテストの分割を使用し、カナリア文字列を尊重する。モデル評価の個々の部分（プロンプトやセーフガードの強さなど）を変化させることも含め、異なる条件や環境のもとで 、複数回モデル評価を再実行する。

外部妥当性確認は、モデルの評価結果が、評価環境外の文脈におけるモデルの挙動の代理とし て使用できることを保証するものである。外部妥当性は、例えば、評価プロセスにおけるドメインエキスパートの適切な統合、適切なケイパビリティの抽出（施策 10.4 参照）、評価が実行される環境条件と実世界のコンテキストから乖離する方法の文書化、適切に保持されたテストセットの利用などによって示すことができる。

再現性とは、同じ入力データ、計算方法、規範、評価条件を用いて一貫性のあるモデル評 価結果を得る能力のことであり、他の研究者や技術者がモデル評価結果を妥当性確認、再現、 改善することを可能にする。これは、例えば、適切な量のモデル評価データの安全な公開（常に拡散リスクを考慮する）、モデル評価規範の安全な公開、評価方法、評価環境、計算環境、及びエリキューション方法の文書化によって促進される、他の当事者によるピアレビューや再現の成功によって示される。

移植性とは、例えば、適切な API や評価標準（特にオープンソース）の上に構築すること、可能な限りモデルにとらわれない評価を維持すること、プライバシーを保護する方法で（例えば、モデルに関する追加情報を他者に漏らすことなく）モデル評価の実装を容易にすることなどを通じて、他の研究者や技術者が一貫してシームレスにモデル評価を実装し、アセスメントできる能力を指す。

さらに署名者は、結果の不確実性のレベルと、使用した手法の限界を報告することを約束する。最終的には、これらの手段を組み合わせることで、署名者は、機械学習や自然科学・社会科学における科学的査読の品質標準を遵守し、最も深刻なリスクに関する重要な結果を、主要な機械学習学会や学術誌への投稿と同様に徹底的に査読することを目指す。

主要業績評価指標の可能性

- **KPI 10.3.1：モデル報告書の証拠として使用されたモデル評価のうち、高い内部妥当性および外部妥当性を示すものの割合（目標：75%）。**
- **KPI 10.3.2：モデルレポートにおいて最高リスク層の証拠として使用されたモデル評価のうち、再現されたものの割合（目標：50%）。**
- **KPI 10.3.3： モデル報告書の証拠として使用されたモデル評価のうち、完全に移植可能なものの割合（非 中小企業の目標：25%、中小企業の目標：0%）。**
- **KPI 10.3.4： 不確実性が報告されているモデル報告書の証拠として使用されているモデル評価の割合（目標：100%）。**

施策 10.4. モデル抽出

署名者は、システミックリスクを伴う汎用 AI モデルの評価において、評価対象モデルの能力、傾向、効果を適切に引き出し、過小評価のリスクを最小化し、潜在的な悪用行為者の現実的な引き出し能力に合致させるため、最新レベルのモデル引き出しを確実に実施することを約束する。これには、例えば、微調整、プロンプトエンジニアリング、（ツールの使用を含む）足場固め、ロジットアクセスの使用、セーフガードを無効にする機能、ベースモデルや有用なモデルのみの使用などが含まれる。

最新レベルのモデルエリシテーションとは、使用されたエリシテーション手法の組み合わせが、評価対象モ デル、または同等のモデルについて最もよく知られた手法の組み合わせと一致し、内部で使用された手法 と、外部で文書化され再現可能な手法の両方の長所を考慮したものである。署名者は、戦略的モデルが評価で不合格となるリスク（サンドバッグや戦略的欺瞞など）を最小化する手法 の使用に努めることを約束する。

そのためには、評価者に十分な計算機予算を提供し、十分な長さの評価実行、並列実行、再実行を可能にする必要がある。また、適切な人員配置とエンジニアリング予算を提供し、例えば、人為的に推定値を下げる可能性のあるソフトウェアのバグやモデルの拒否を特定するために、結果を精査する必要がある。最も深刻なリスクが特定された場合、署名者は、安全性以外の主要研究プロジェクトに費やされるエリシテーショ ンの労力とほぼ同額にすることを約束する。リスクレベルが低い場合は、これらの労力を比例的に削減することができる。

加えて、署名者は、リスクプロファイルが増加する（例えば、モデルを有害な能力により適したものにする、あるいは有害な傾向を増加させる）エリシテーション手法については、核拡散リスクやエリシテーションされたモデルによる有害な行動を防止するため、セキュリティ対策（対策 12.2 参照）を強化することを約束する。

主要業績評価指標の可能性

- **KPI 10.4.1：最高リスク層の証拠として使用されるモデル評価のために、モデル抽出に費やされたエンジニアリング時間の、社内の最大規模の非安全プロジェクトと比較した平均割合（目標：75%）。**
- **KPI 10.4.2：最大規模の社内の非安全プロジェクト（目標：33%）と比較して、低リスク層の証拠として使用されるモデル評価のために、モデル抽出に費やされたエンジニアリング時間の平均割合。**

施策 10.5. システムの一部としてのモデル

署名者は、システミックリスクを持つ汎用 AI モデルを AI システムに組み込んだ場合、モデルアセスメントでその能力と限界を確実に評価できるようにする。システミックリスクを持つ汎用 AI モデルを AI システムに使用または統合する予定の署名者は、評価対象のリスクに関連する場合、モデル評価において将来の展開状況を考慮し、それに応じてモデルを評価する（例えば、 同様の足場やツールを使用して評価する）。展開環境が当初アセスメントに使用したものから大きく変化する場合は、これを考慮し、6.3 項で定義するとおり、（部分的または完全な）モデルの再評価を実施する。

(例えば、オープンソースリリースや B2B ライセンスなど) 署名者自身が AI システムで汎用 AI モデルを使用または統合しない場合、ライセンシーは、ライセンシーが使用または展開を予定している AI システムを代表する AI システム内でモデルを評価するよう、ライセンス条件を実施する。この評価は、特に最高リスク層のアセスメントや、ライセンシーのリソースを考慮し、シグナト リオがこのモデルに適用しているのと同じ標準の厳密さで実施される。加 盟者は、このモデル評価に必要な科学的情報およびツールをライセンシーに提供することができる（方策 10.3 お よび 10.8 参照）。

施策 10.6. 代表者モデルの評価と一般化

署名者は、モデルの評価がモデルの想定される使用状況に合致することを保証する。例えば、多言語モデルの言語ベースの評価では、英語だけでなく、リスクアセスメントに関連する限り、 主要な欧州言語（または、モデルがサポートすると主張するその他の言語）も考慮する（リスクによっては、 英語で評価するのが最適と思われる能力、すなわちプログラミング能力に基づくものもある）。署名者は、テキストだけでなく、モデルがサポートするすべてのモダリティについて、このようにすることを約束する。他のモダリティについては、言語とは異なるカテゴリーで表現力を評価する（画像表現など）。

これを促進するため、署名者は、リスクアセスメントに適切であれば、市民社会、学術界、その他関 連ステークホルダーの専門家または一般代表者が参加する参加型手法を利用することが奨励される。関連する利害関係者（例えば影響を受ける人々）がいない場合、署名者は、彼らの利益を代表できる最良の 代表者を特定し、証拠収集プロセスにおいて彼らと協議することを約束する。

施策 10.7. 探索的作業

調印者は、システミック・リスクを伴う汎用 AI モデルについて、研究機関の規模やアセスメントするリスクに見合った探索的作業、例えば探索的リサーチやオープンエンドのレッドチーム（特に市民社会やその他の影響を受けるステークホルダーの視点を含む）を確実に実施することを約束する。また、既知の能力に関する証拠収集だけにとどまらず、これらの手法を通じて、新たな能力、傾向、影響によるリスクのアセスメントにも努めることを約束する。

主要業績評価指標の可能性

- **KPI 10.7.1：中小企業以外の目標：10%、中小企業および／またはオープンソースモデルを独占的にリリースしているプロバイダの目標：0%).**

施策 10.8. ツールとベストプラクティスを共有する

署名者は、最先端の証拠収集の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にすることを約束する。また、特に他のモデルプロバイダーやサードパーティ評価者のために、適切なモデル評価ツールを提供する。特に、署名者がすでに社内で多大なツールや経験を培っている場合は、その専門知識を中小企業や新 規プロバイダと共有することを目指す。

しかし、拡散リスクの高いモデル評価データ（すなわち、テストセットが悪用行為者の訓練セットになる可能性がある）や科学的厳密性を脅かすようなモデル評価データ（すなわち、テストデータを訓練セットに流出させたり、現在アクティブなまま保持しているテストセットを完全に公開したりする）には、さらに注意を払う必要がある。適切な場合、これらの考慮事項はモデル報告書に記載される。また署名者は、公共の安全や商業上の機密情報を保護するため、安全上の利益を上回る情報の共有を制限することができる。

署名者は、上記を共有することで、そもそも安全研究を行う能力が犠牲にならないようにすることを約束する。既存の安全研究者の仕事量はそのままに、この作業を促進するために、研究チームに技術スタッフやサポートスタッフを追加雇用または配置することを検討する。

このプロセスを AI 事務局または AI 事務局が指定するその他の公的機関（例：AI 安全性機関の国際ネットワーク）が促進しない場合、署名者は業界団体を通じてこの共有を促進することが奨励されるが、潜在的な競争相手の排除につながる厳格な会員資格要件やその他の制限がないことを保証する。

主要業績評価指標の可能性

- **KPI 10.8.1：証拠収集のベストプラクティスガイダンスを共有した関係者の数（中小企業以外の目標：10、中小企業の目標：0）。**
- **KPI 10.8.2：少なくとも 5 つの他のプロバイダまたは評価者と共有された新しいモデル評価ツールの数 （中小企業以外の目標：3、中小企業およびオープンソースモデルを実質的に公開しているプロバイダの目標）：0).**

施策 10.9.有資格のモデル評価者と評価へのアクセス

署名者は、モデル評価チームが業務に必要な資格を有することを保証する。これには、機械学習の能力だけでなく、リスクやその他の関連事項を効果的に評価するために必要な、関連分野の専門知識へのアクセスも必要となる。

署名者は、すべてのモデル評価者（内部および/または外部）に対し、システミックリスクを伴う汎用の AI モデルを適切に評価するのに十分な時間、モデルへのアクセス、エンジニアリングサポート、計算予算を提供し、厳密な科学的標準（評価基準 10.3 に記載）に沿った作業に必要なサポートを提供することを約束する。これには、適切なレベルのモデルエリシテーション（施策 10.4 参照）や、適切な場合にはシステムの一部としてのモデルへのアクセス（施策 10.5 参照）も含まれる。さらに、外部のモデル評価者については、施策 16.1 に定める条件が適用される。

モデル評価を容易にするため、署名者は、モデルセキュリティに関する費用便益分析を考慮した上で、システミックリスクを含む汎用 AI モデルに、より多くのモデル評価者（社内外）、グレーボックス、ホワイトボックスからアクセスできるようにすることを約束する

（施策 12.2 参照）。

施策 10.10. 安全マージン

リスク階層と緩和策の対応付けにモデル評価結果を利用する場合（評価 4.3 項参照）、署名者は、潜在的なエリキュー ション不足、展開後の潜在的改善、証拠収集における一般的不確実性を代表する、かなりの安全マージンを定義 することを約束する。

施策 10.11. 予測対応モデルの

例えば、将来のモデルがいつ危険な能力を示すかを予測する「スケーリング則実験」を行うなどである。

主要業績評価指標の可能性

- **KPI 10.11.1：モデルレポートの証拠として使用された評価のうち、ある程度の予測を可能にするものの割合（目標：10%）。**

施策 10.12. 展開後のモニタリング

署名者は、最高リスク段階の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ことを確約し、低リスク段 階についても同様に実施することができる。これには、第 10.1 項で定義される、関連するモデルから独立した証拠の収集が含まれるが、これに限定され るものではない。

署名者は、リスクアセスメントに展開後の関連情報を収集し、含めるための契約上、技術上、またはその他の 仕組みを確立し、その緩和策の有効性を確認する。展開後の関連情報とは、モデルを上市した後に署名者が収集できる情報を指し、リスクと緩和策の特定と評価に役 立つ。このような情報には、モデル統合、モデル利用、モデルへの影響に関する情報が含まれるが、これらに限定されるものではない。これらの仕組みは、モデルの統合や使用方法によって異なる可能性があるため、署名者は、モデルの流通戦 略や、モデルを使用する顧客や業種に合わせて、展開後の監視を行うことが推奨される。

すべての署名者（オープンソースモデルのプロバイダを含む）は、以下のような方法を検討することが推奨される：匿名での報告チャンネル、インシデント報告フォーム、バグ報奨金、 コミュニティ主導の評価や公開リーダーボードのサポート、現実世界におけるモデル利用の証拠を監視する（例えば、ソフトウェアリポジトリや既知のマルウェアにおける利用を特定したり、パブリックフォーラムやソーシャルメディアを監視して新しい利用パターンを見つける）；モデル・エージェントから構築された個々の AI インスタンスのプライバシーを保護したモニタリングと分析をサポートする新しい技術的手法に投資する（例えば、電子透かし、フィンガープリンティング、デジタル署名、分散型のプライバシー保護モニタリング技術など）。

API 経由で展開されるクローズドソースモデルのプロバイダである署名者は、システミックリスクのある汎用 AI モデルを、軽量なテレメトリー（ロギング）とデータ分析手法によって追加的に監視し、EU の法律や規制に従って、ユーザーのプライバシーを確実に保護し、機密データの送信を回避することが奨励される。モデルプロバイダー自身が AI システムを展開する場合は、特にエンドユーザーのフィードバックを考慮し、これらのシステムの一部としてこれらのモデルを監視する。

加えて、特に上記の方法が利用できず、署名者が他の手段で十分な関連情報を入手できない場合、署名者は、川下 のモデル展開者や利用者と、署名者が調査したシステミック・リスクに関する重要な証拠を、プライバシーを保護する形で 共有することを義務付けるライセンス契約を利用し、それを実施する。自然人であるエンドユーザーに対しては、署名者は、エンドユーザーが同様の関連情報を共有することをオ プトインすることを許可し、EU 法および規制に従い、ユーザーのプライバシーを保護し、機密デー

タの送信を回避する。

主要業績評価指標の可能性

- **KPI 10.12.1：展開後モニタリングの少なくとも 3 つの方法を用いて証拠を収集している最高リスク段階のリスクの割合（目標：75%）。**
- **KPI 10.12.2：展開後のモニタリングの少なくとも一つの方法を用いて証拠を収集している、低リスク層のリスクの割合（目標：50%）。**
- **KPI 10.12.3 最高リスク段階について収集された実施可能な証拠を報告するライセンシーの割合（目標：50%）。**

コミットメント 10（証拠収集）の変更点の説明：このセクションのほとんどの変更は、特定の重要な用語（「科学的厳密性」や「最新のモデル評価」など）の定義方法を明確にするよう求めたことに基づくものである。

また、「施策」を満たすために利用可能な選択肢が、目の前のリスクに比例するようにすることも目指した（特に、リスクの最も高い階層にさらに焦点を当てることができるようにした）。さらに、展開の状況（たとえば、オープンソースのリリースにはどのような方法が適切か）を考慮し、中小企業の官僚的な負担を減らすと同時に、可能な限り明確化することを目指した。また、このコミットメントは、内部または外部（ ）で行われる仕事に適用できることを明確にした（コミットメント 16 も参照）。

また、（ここに移された）展開後のモニタリングは、固定的で柔軟性に欠け、不適切な可能性のある単一の手法として規定されたものではなく、ビジネスモデルやモデル展開の状況に応じて最も適切な方法で実施可能な施策として規定されたものであることを明確にしたかった。

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

コミットメント 11. 安全緩和

AI 法第 55 条(1)項(b)に基づく義務を果たすため、署名者は、そのようなモデルの開発・展開によるリスクに見合った安全緩和策を実施し、リスクを許容可能なレベルまで軽減することを約束する。署名者は、安全緩和策が最新のものであることを保証し、利用可能な場合には AI 事務局のガイダンスに従う。

コミットメント *11* を達成するために、以下を行う：

施策 11.1. 考慮すべき安全緩和策

署名者は、以下を含むがこれに限定されない安全緩和策の実施を検討するよう奨励される：(a)訓練データのフィルタリングとクリーニング、(b)モデルの入力と出力の監視とフィルタリング、(c)特定の要求を拒否するためのモデルの微調整など、モデルの動作の変更、(d)特定のモデルへのアクセスを審査されたユーザに制限するなど、モデルの展開の制限、(e)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やその他の安全ツール、(f)モデルの動作に関する高保証の定量的安全保証。

施策 11.2. 最先端技術

最先端の緩和策とは、研究、技術、経験から得られた知見を最もよく反映し、システミックリスクを緩和 するための最もよく知られた手順を取り入れたものである。最先端の緩和策とは、必ずしも全てのリスクを最大限に緩和するものではなく、署名者はリスクに見合った 緩和策を実施するよう努力する（4.2 項のリスク階層参照）。この意味で、最先端の緩和策とは、特に高次のリスクを最も効果的に緩和するものである。また、緩和策の実施によって失われる利益も考慮することが奨励される。

コミットメント 11 の変更点の説明：これまでは、安全とセキュリティの緩和は、単一の「施策 12」でカバーされていたが、このドラフトでは、コミットメント 11 と 12 に分割されている。

回答者の中には、どのような安全緩和が必要なのか、より詳細な説明が必要だと考える者もいた。 、技術的リスク緩和の分野は発展途上であり、有効であることが確認されている確立された解決策がほとんどないという事実を考慮しつつ、より詳細な情報を提供するため、署名者が検討することが推奨される安全緩和策を列挙する。

我々は、アウトカムベースのアプローチが最も効果的であると考える。将来、時代遅れになるかもしれない（あるいはもっと悪いことに、積極的に逆効果になるかもしれない）固定的な緩和策のセットを要求する代わりに、モデルプロバイダーは、モデルがもたらすリスクに比例した緩和策の実施を約束すると考える。

コミットメント 12. セキュリティ緩和

AI 法第 55 条 1 項(b)および(d)の義務を果たすため、署名者は、システミック・リスクを伴う汎用 AI モデルの未公開の重み、および未公開のアルゴリズム洞察などの関連資産を保有することによるシステミック・リスクを緩和する。そのために署名者は、ランド研究所 SL3 ベンチマークに沿って、インサイダー脅威を含む、高い動機と十分な資金を持つ非国家アクターが、そのようなウェイトや関連資産を入手しようとする試みを阻止するためのセキュリティ緩和策を実施する。署名者は、国家レベルの敵対者からそのような分銅や関連資産を盗み出そうとする試みに直面することが予想される場合、ランド研究所 SL4 以上のベンチマークに沿った、より厳格なセキュリティ緩和策を実施し、研究を進めることが奨励される。

コミットメント *12* を達成するために、以下を行う：

施策 12.1. 一般的なサイバーセキュリティのベストプラクティス

署名者は、一般的なサイバーセキュリティのベストプラクティスを採用することを約束する。これには以下が含まれるが、これらに限定されるものではない：

(a) 強力なパスワード・ポリシーと管理を行う、

(b) 不審なリンクやフィッシング詐欺のメールをフィルタリングする、

(c) ワイヤレスネットワークの保護、

(d) 信頼できないリムーバブルメディアに対するポリシー、

(e)物理的な侵入防止、定期的なソフトウェア更新とパッチ管理。

これらのベストプラクティスは、ISO/IEC 27001、NIST 800-53、SOC 2、もしあればサイバーセキュリティに関する欧州統一標準などのサイバーセキュリティ技術標準に沿ったものであり、NIS2 指令やサイバーレジリエンス法などの規制にも沿ったものとなる。

施策 12.2. セキュリティ保証

署名者は、「約束 12」の脅威アクターに対するセキュリティ態勢を保証し、テストするための施策を実施することを約束する。これには以下が含まれるが、これに限定されるものではない：

(a) 認定サードパーティによる頻繁なセキュリティ・レビュー、

(b) レッドチームが頻繁に活躍している、

(c) サードパーティがセキュリティ問題を報告するためのセキュアなコミュニケーションチャネルを確保する、

(d) 競争的なバグ報奨金プログラムを実施し、セキュリティテストへの一般参加を促す、

(e) 報復を禁止する明確かつ公安的な内部告発ポリシーは、以下の通りである。

コミットメント 18

(f) エンドポイント検知・応答（EDR）ツールと侵入検知システム（IDS）ツールを社内のすべてのデバイスとネットワークコンポーネントに導入し、セキュリティチームを使用して EDR アラートを監視し、セキュリティ侵害に対するインシデントハンドリング、対応、復旧をタイムリーに実行する。

これらの対策は、NIST800-53（CA-2(1)、CA8(2)、RA-5(11)、IR-4(14)）や NIST SP 800-115 5.2 などのセキュリティ保証技術標準と整合する。

施策 12.3. 保管されているモデルウェイトおよび関連資産の保護

署名者は、保存されたモデルウェイトと、未発表のアルゴリズム的洞察などの関連資産を保護するため の対策を実施する。これには以下が含まれるが、これに限定されるものではない：

(a)モデルウェイトが保存されているすべてのデバイスと場所のレジストリ、

(b)モデル・ウエイトを保存するすべてのデバイスのアクセス管理およびアクセス監視を行い、管理されていないデバイスへのコピーには警告を発する、

(c)データ、規範、および同等の機密性とセキュリティレベルで扱われるサービスのみをホストする専用デバイス上のストレージである、

(e)モデル分銅が合法的に使用される場合にのみ、非永続的なメモリに復号化されることを保証する、(g)データセンター及びその他の機密性の高い作業環境への物理的なアクセスを、必要な人員のみに制限し、そのような場所に権限のない人員や装置がないか定期的に検査する。

これらの対策は、NIST 800-53 などの技術標準に沿ったものとなる。

施策 12.4. モデルウェイトへのインターフェースとアクセス管理

署名者は、保管中または使用中のモデル重量のインターフェースやアクセス方法を強化する施策を実施する。これには以下が含まれるが、これらに限定されるものではない：

(a) モデルウェイトへの直接的または間接的なアクセスについて、必要なソフトウェアと要員の事業体のみに 明示的な権限を与え、多要素認証メカニズムによって強制し、定期的に監査する

(b) モデルウェイトにアクセスするソフトウェアインターフェイスの脆弱性やデータ漏洩について、セキュリティチームによる徹底的なレビューを

(c) データおよびウェイトが流出するリスクを低減するため、ウェイトにアクセスするインターフェイスをハード化する（出力レート制限などの方法を使用）。

これらの対策は、NIST SP 800-171、INCITS 359-2004、NIST 800-53 などの技術標準に沿ったものとなる。

主要業績評価指標の可能性

- **KPI 12.4.1：未リリースのモデルウェイト及び関連資産に対す る、定義されたインタフェース及びアクセスポ イントのうち、指定されたセキュリティ管理が実施されたものの割合（目標：100%）。**

施策 12.5. インサイダーの脅威

署名者は、インサイダーの脅威を選別し、保護するための対策を実施することを約束する。これには以下が含まれるが、これらに限定されるものではない：

(a) 未発売のモデル分銅、関連資産、又は未発売のモデル分銅の保管又は使用を制御するシステムへのアクセスを有する、又は合理的に得る可能性のある従業員及び請負業者の身元調査、並びに内部脅威の認識及び報告に関する研修の実施。

これらの対策は、NIST800-53（PM-12、PS-3）のような技術標準に沿ったものとなる。

施策 12.6. 適用範囲

署名者は、訓練前からモデル重量の安全な削除または公開までのライフサイクル全体を通じて、システミックリスクを伴う汎用 AI モデルの未公開モデル重みおよび関連資産のすべてのバージョンおよびコピーに、第 12.1-12.5 項のセキュリティ対策を確実に適用することを約束する。

主要業績評価指標の可能性

- **KPI 12.6.1：対策 12.1.～12.5.のセキュリティ緩和策を適用した、未リリースのモデルウェイトのコピーとバージョンの割合（目標：100%）。**

施策 12.7. 限定的開示

署名者は、「枠組み」または「モデル報告書」の一般に入手可能なコピーにおいて、「施策 12.1-12.5」のセキュリティ緩和策を、セキュリティ緩和策の有効性を損なわず、かかる緩和策に関する一般に入手可能な詳細と矛盾しない程度にのみ開示することを確約する。

主要業績評価指標の可能性

- **KPI 12.6.1：対策 12.1.～12.5.のセキュリティ緩和策を適用した、未リリースのモデルウェイトのコピー及び バージョンの割合（目標：100）**

コミットメント変更点の説明：従来、安全性とセキュリティの緩和は、一つの「施策 12」でカバーされていたが、本ドラフトでは、コミットメント 11 と 12 に分割されている。緩和策をより具体的なものにするというフィードバックに応え、我々は、 、 署名者にとって適切と思われるセキュリティ緩和のレベルを明記した。このセキュリティ緩和レベルに関するフィードバックを期待する。

コミットメント 13. 開発と展開の決定

AI 法第 55 条 1 項(b)および(d)の義務を果たすため、署名者は、安全・セキュリティの緩和が不十分な場合、残存リスクを緩和することを約束する。そのために署名者は、システミック・リスクを持つ汎用 AI モデルの開発・展開を進めるかどうか、またどのように進めるかを決定するプロセスを確立する。このプロセスは枠組みに記載され、このプロセスの決定方法と、管理組織を含む組織内でのエスカレーション方法を含む（コミットメント 14 参照）。

コミットメント *13* を達成するために、以下を行う：

施策 13.1. 手続きを進めない条件

(a)リスクを許容できないレベル以下に抑制する、または、(b)許容できないレベル以下のリスクを適切に緩和することが不十分であるため、システミック・リスクを伴う汎用 AI モデルの開発・展開を進めない条件を、枠組み内に詳述することを署名者に約束する（第 4.2 項参照）。また、展開後のモニタリングとリスクアセスメントに基づき、必要であれば、（公 開されていない）モデルをデデプロイし、別のモデルにフォールバックするためのプロセスも記述する。このコミットメントは、開発と展開の意思決定に用いられる他の条件やプロセスの使用を妨げるものではない。

施策 13.2. 進行の決定プロセス

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・展開を進めることができないとの決定がなされた後、その決定プロセスを枠組みに詳述することを約束する（第 13.1 項参照）。これには、適切な安全・セキュリティ緩和策の実施と、システミックリスクが許容できないレベル以下に抑えられ、そのレベル以下に適切に緩和されていることを証明するために必要なリスクアセスメントが含まれる。

施策 13.3. 進行中の展開ステージ

署名者は、システミックリスクを伴う汎用 AI モデルの展開を、システミックリスクを許容できないレベル以下に抑え、それ以下では適切に緩和するために、どのようなステージを設けるのか、またどのように展開するのかを枠組みで詳述することを約束する。これには、(a)API へのアクセスを吟味されたユーザーに限定する、(b)リスクアセスメントに基づいて段階的にアクセスを拡大する、(c)オープンリリースの前にクローズドリリースから開始する、(d)ロギングシステムを使って利用状況や安全性の懸念を追跡する、(e)リスク階層やユーザーからのフィードバックに基づくものを含め、段階的に進めるための規準を設定する、(f)アクセスを制限する能力を保持する、といった慣行が含まれる。

施策 13.4. 意思決定における外部意見の透明性

容認できないリスクのマネジメントに関する意思決定に透明性を持たせるため、 署名者は、特に容認でき ないリスクが関係する

場合、開発および展開の決定に、関連する政府関係者を含む外部関係者の意見を取り入れるかど うか、いつ取り入れるかについて、その枠組みに詳述することを約束する。このコミットメントを達成するためには、そのような関与や認可は必要ないが、フロンティア AI の安全性に関するコミットメント（成果 3.VIII）など、既存の国際的なコミットメントに沿った形で、その有無に関する透明性を確保する必要がある。

コミットメント 13 の変更に関する説明：従来、コミットメント 13 は「施策 14」と付されていた。

すなわち、(a) 進行を決定する際の展開ステージの記述など、意思決定プロセスの記述に含めるべき内容について、より詳細な説明と例を示すこと（新たな Measure 13.3）、(b) このコミットメントでは、外部からの意思決定へのインプットは必須ではなく、そのようなインプットが必要な場合にのみ記述を求めることを明確にすること（Measure 13.4）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対するガバナンス・リスク緩和

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、AI 法第 55 条(1)項(b)に従い、連邦レベルで起こりうるシステミック・リスクを緩和することが求められる。本項では、この義務を遵守するために、署名者が約束するガバナンス施策とプロセスを定める。またこの章では、第 52 条(1)、53 条(1)の(a)項（特に附属書 XI、AI 法第 2 項）、第 55 条(1)の(c)項、AI 法第 87 条の義務を履行するために、 署名者が約束する施策についても定める。

コミットメント 14. システミック・リスクの責任分担

AI 法第 55 条 1 項の義務を果たすため、署名者は、汎用 AI モデルから生じるシステミックリスクをアセスメントし、比例的に緩和する義務を果たすことを視野に入れ、マネジメント機関（マネジメントと監督機能の両方）、商品、開発、その他の関連業務チームなど、組織全体に適切なレベルの責任とリソースを明確に割り当てることを約束する。

これには、組織の複雑性、ガバナンス構造、システミック・リスクを伴う汎用モデルがもたらす可能性のあるシステミック・リスクに応じて、適切なレベルの責任とリソースを割り当てることが含まれる。コミットメント 14 には、システミック・リスクを生み出す可能性のある中核的な事業活動に重点を置く組織の部分、組織のシステミック・リスクの評価と緩和を監督、実施、支援する部分、リスク評価と緩和の適切性について独立した保証を提供する部分の間で、明確かつ適切な責任を割り当てることが含まれる。

署名者は、健全なリスク文化を推進することが奨励される。これには、トップからの基調設定、効果的なコミュニケー ションと挑戦の許可、過度なリスクテイクを抑制するための適切なインセンティブ（慎重な行動に対する報奨やリ スクの社内への報告など）が含まれる。

主要業績評価指標の可能性

- **KPI 14.1：**
    - **中小企業の場合：エグゼクティブ・チームの特定の個人が、組織のリスクマネジメントの取り組みの責任を割り当てられている。この個人は、その職務を遂行するために十分な人員、時間、支援を得ている。**
    - **非中小企業の場合：スリーラインズモデル（内部監査人協会、2024 年など参照）に沿って、署名者は、署名者のシステミック・リスクマネジメントの取り組みを支援・監督するチーフ・リ スク・オフィサー（または同等者）を任命している。チーフ・リスク・オフィサーは、十分に人員配置されたリスク機能によりサポートされている。**
- **KPI 14.2：**
    - **リスク委員会や監査委員会などの特定の個人に、システミック・リスクのマネジメントを監督する責任と適切なリソースが与えられている。これは、関連する個人に提供される外部監査報告書を通じて達成することができる。**
    - **監督機能を有するマネジメント組織、またはその適切なサブセットは、最低限、署名者が文書化したすべての情報を含め、組織のシステミックリスクマネジメントを監督するために必要な情報にアクセスできる。**

- **中小企業の場合：監督機能を有するマネジメント組織のサブセットが、組織のリスクマネジメントの取り組みを監督する責任を割り当てられ、その任務を遂行するための十分な資源が提供されている。**
- **非中小企業の場合：スリーラインモデル（内部監査人協会、2024 年など参照）に沿って、マネジメント組織の監督機能のサブセットが、例えば監査委員会やリスク委員会を設置するなどして、組織のシステミック・リスク管理の取組みを監督する責任を割り当てられ、十分なリソースを確保されている。このような組織は、組織のコンプライアンスとリスクマネジメントの取り組みについて、独立した証拠を提供するために、直接報告する内部監査機能によってサポートされる。**

- **KPI 14.3：コミットメント 15 に従った）堅持と適切性のアセスメントにより、システミック・リスクを評価し、比例的に緩和するために、適切なレベルの責任と資源が組織全体に割り当てられていると結論づけられる。**

コミットメント 14 の変更点の説明：ステークホルダーからのフィードバックに応え、責任が様々なレベルにまたがる可能性があるという事実を認識し、様々な組織構造に対応できるよう文言が改善された。デジタルサービス法第 41 条など、確立された用法に沿った「管理団体」という用語が採用された。枠組みを、取締役会と経営陣の機能が分かれていないものを含む、さまざまなガバナンスモデルに適応できるようにするため、枠組みは一つのコミットメントに統合された。加えて、内部監査機能や健全なリスク文化の醸成など、強固な監督手段に対するステークホルダーからの提案に対応するため、新たな KPI と自主的施策が導入された。

コミットメント 15．枠組みの順守とアセスメント

AI 法第 55 条 1 項の義務を果たすため、署名者は、約束 4 で採択・実施を約束した枠組みの遵守と適切性を評価する。

署名者は、システミック・リスクを伴う汎用 AI モデルについて、上市後 6 ヶ月以内にリスクアセスメントを実施する。署名者は、AI 法第 51 条の規準に合致している、または合致する予定であることを AI 事務局に通知してから 4 週間以内、または 6 ヶ月ごとのいずれか早い時期に、自国の枠組みのアセスメントを実施することを約束する。例えば、リスク状況に大きな変化があった場合や、システミック・リスクを伴う汎用 AI モデル（ ）が重大な役割を果たしたとみなされるような重大なインシデントが発生した場合など、署名者はより頻繁にこのようなレビューを実施することが奨励される。

これらのアセスメントは、（該当する場合は内部監査機能などを通じて）加 盟者自身の職員が実施することも、（16.1 項に従って）外部機関が実施することもある。順守とアセスメントの両方が、管理団体の監督機能において報告される。レビューの徹底度は、当該モデルがもたらす可能性のあるシステミック・リスクに比例する。

アセスメントは、ベストプラクティス、関連研究、最先端科学、重大インシデント報告、外部の専門家の可能性を考慮する。十分な証拠に裏付けされた以下の質問を対象とする：

- 枠組みは、その有効性を評価するのに十分な詳細さで、コミットメント 4 に概説されているすべての構成要素に対応しているか。その他、枠組みは実務規範と整合しているか。
- 枠組みは、リスクアセスメントや緩和手法の有効性だけでなく、将来の開発における不確実性を適切に考慮しているか。
- 今後 9 ヶ月間、枠組みが遵守されると信じる強い理由はあるか？例えば、この枠組みは、この期間内に上市される予定のモデルに対して適切であると思われるか？

- 今後 9 ヶ月以内に上市が予定されている組織のモデルがもたらすシステミックリスクの複雑さと規模に、枠組みの方針、リソース、緩和策が依然として比例しているか。

アセスメントは、モデルを市場に出してから 6 カ月後に実施し、十分な証拠に裏付けられた以下の質問を取り上げる：

- システミック・リスクを伴う汎用 AI モデルは、関連する枠組みを遵守していたか？
- 関連するモデル報告書は枠組みを遵守していたか。モデル報告書は、枠組みへの準拠を評価するのに十分な証拠を提供したか。
- 市販後の緩和とアセスメントは、モデル報告書に従って実施されたか？
- これまで枠組みで識別されていなかった新たなリスクがモデル報告書に記載されたか。

主要業績評価指標の可能性

- **KPI 15.1：アセスメントを 6 カ月ごとに実施する。**
- **KPI 15.2：アセスメントとアドヒアランス評価で肯定的な結果が得られた。**
- **KPI 15.3：適合性評価とアドヒアランス評価で識別された懸念事項は、当該モデルが上市される前、または次回予定され ている適合性評価とアドヒアランス評価のいずれか早い時期までに対処する。**
- **KPI 15.4：外部監査人が、遵守または適切性のアセスメントを実施し、肯定的な結果を得る。**
- **KPI 15.5：枠組みは、類似のリスクプロファイルを有するシステミックリスクを有する汎用 AI モデルの他のプロバイダが使用する枠組みに比べ、厳密性、幅、深さにおいて遜色がない。**

コミットメント 15 の変更点の説明：ステークホルダーからのフィードバックに応え、遵守と適切性のアセスメントに関する規定を拡大し、その範囲、方法、時期を明確にした。これらの調整は、枠組みやモデル報告書について新たな標準を設定することを意図したものではなく、むしろ、既存の枠組みや報告書が規範との整合性を維持し、それらの継続的な整合性を裏付ける証拠が収集されていることを確認するためのものである。アセスメントの目的は、署名者が現在の実務がどのような点で 不十分であるかを特定し、モデル作成中に調整する時間を与えることである。

コミットメント 16. 外部リスクアセスメント

AI 法第 55 条 1 項に基づく義務を果たすため、署名者は、コミットメント 6-12 に従い、システミックリスクを有する汎用 AI モデルから生じうるシステミックリスクを評価し緩和する義務を果たすべく、必要に応じて、そのライフサイクルを通じて、外部の専門家による十分なリスクアセスメントと緩和アセスメントを可能にすることを約束する。

コミットメント *16* を達成するために、以下を行う：

施策 16.1. 市場投入前

署名者は、システミックリスクを伴う汎用 AI モデルについて、市場導入前（第 6.2 項および第 6.3 項による） に、コミットメント 6-10 項に従い、有資格の外部アセスメント担当者がリスクアセスメントの関連部分を実施することを確約する。このコミットメントは、コミットメント 7 に従って特定されたシステミック・リスクのサブセットに限定される：

- プロバイダが、そのモデルが、例えば斬新なモデル機能や実装された緩和策の違いにより、EU 市場に既に存在するシステミックリスクを有する汎用 AI モデル以上の追加的なリスクをもたらさないという十分な証拠を提示できない（コミットメント 10 で定義される EU 市場に既に存在するモデルとの類似性は、そのような証拠となる）；

  または

- 署名者は、リスクアセスメントを効果的に実施するために、関連する社内の専門知識または情報が不十 分である（国家安全保障に関連する機密情報を必要とするアセスメントなど）。

上記の規準が満たされているかどうかの説明を、モデル報告書に記載する。上記の規準が満たされているにもかかわらず、署名者が誠意をもって努力した結果、下記の規準に従っ て適格な外部アセスメント実施者を特定できなかった場合（早期の捜索努力、適格なアセスメント実施者が見つか った場合は、下位施策 6.1 に従って外部評価者に早期に通知することを含む）、、内部評価 に頼り、外部評価の欠如を正当化する理由をモデル報告書に記載することを確約する。

特に、リスクアセスメントと緩和の取り組みに関して、ベストプラクティスとの整合性を評価する。

署名者は、モデルの上市を決定する際、可能であれば外部アセスメントで収集した情報を考慮し、モデル報 告書にそのような評価を含める。外部評価者の勧告や結論に同意できない場合、署名者は、モデル報告書にその旨を明記する。

アセスメントは、AI 事務局、または AI 事務局が認めた、もしくは以下の規準を満たす、政府団体を含む有資格の外部評価者によって実施される：

- 評価対象とするリスク領域またはリスクモデルについて、その領域に関する専門知識を有す る、
- 該当する場合は、技術的なスキルがあり、モデル評価の実施経験がある、
- 付与されるアクセスのレベルと種類に十分な、内部および外部の情報セ キュリティ・プロトコルを備えている、
- 適切な利益相反対策など、結果の公平性を確保する契約条項などのセーフガードを含め、プロバイダから十分なレベルの独立性を有している。

外部評価者を起用する場合、署名者は、上記の規準に基づき、評価者選定の正当性を模範報告書に記載する。該当する場合、外部評価者が AI 事務局により認定されている場合は、正当な理由を記載する必要はない。

外部評価者がモデル評価を実施する場合（コミットメント 10）、署名者は特に、十分なモデルへのアクセス、モ デルに関する情報、時間、資源（施策 10.9）、実験の非保持、独立した結果報告能力、モデルへのアクセス制限を 含む非報復の保証を確保することを約束する。

署名者は、このような外部評価を通じて得られた証拠を利用し、コミットメント 6-10 の証拠を提供する責任を果たすよう 奨励される。

特定のシステムリスクを展開前に外部アセスメントすることを、署名者が確約するのはどのような場合か。

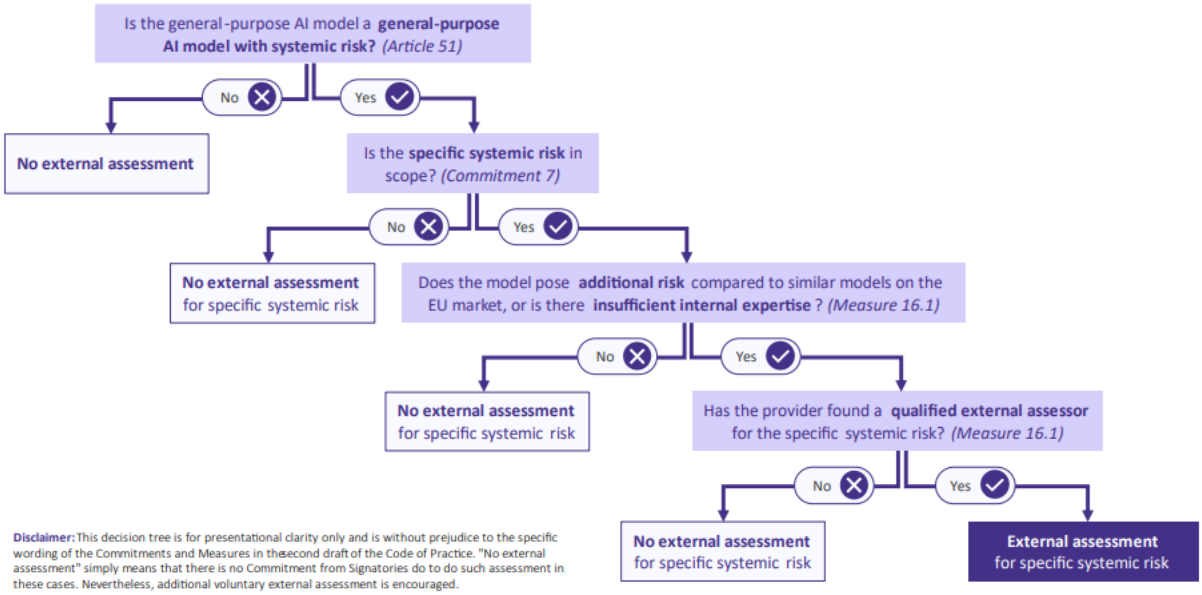

図 *1*：特定のシステミック・リスクについて、展開前に外部アセスメントを実施する条件を示すデシジョンツリー。

施策 16.1 の変更についての説明：この改訂は、複数の利害関係者グループから寄せられたフィー ドバックに直接対応したものであり、最初のドラフトで提案されたも のよりも厳格な外部評価要件を求める意見もあれば、コミットメント は完全に任意であるべきだと主張する意見もあった。すべての利害関係者は、外部審査員に関する「適切性」が何を意味するのかについて、かなり詳細な説明を求めていた。更新された文章では、誰が「資格のある」外部審査員として認定されるのかが明確にされ、AI 事務局が認定する外部審査員のリストなど、継続的に更新されるガイダンスが示されている。

特にプロバイダは、エコシステムがまだ未成熟であることを踏まえ、外部評価の厳格な義務化について懸念を表明し、施策 16.1 が事実上の事前承認体制につながりかねないとの懸念や、そうした外部評価が不必要で負担が大きいのではないかという懸念を表明した。同時に、他の利害関係者は、展開前の十分かつ義務的な外部試験の重要性を引き続き強く強調している。これらの懸念に対処するため、今回の改訂では、具体的にどのような場合に外部アセスメントが適切で、どのような場合にそのような外部アセスメントが不要であるかを明確にした。これは前文 114 の AI 法の解釈に沿ったものである。さらに、外部アセスメントが適切な場合であっても、プロバイダが特定の状況下で、なぜ外部アセスメントを実施できないのかを正当化できるようにすることで、柔軟性を維持している。さらに、外部アセスメントの結果は、（展開前ではなく）展開時にのみ提出されるモデル報告書に記載され、プロバイダは（外部アセスメントが行われた場合）外部アセスメントの勧告に反対する自由があることを保証することで、プロバイダがそれに従う必要がないことを明確にしようとしている。最後に、この文章は、有資格アセスメントの規準を満たし、ガバナンス 6～10 に概説されているリスクアセスメント基準を遵守している限り、国際政府団体と外部アセスメント機関の間で、外部アセスメント結果を相互承認する可能性を開いている。特に、有資格アセスメント機関とみなされるための要件として、EU 居住権（ ）を明記していない。

施策 16.2. 市場投入後

署名者は、システミック・リスクを伴う汎用 AI モデルが、市場導入後にシステミック・リスクを特定、評価、緩和するために、外部アセスメント機関による十分な探索的評価を受けることを保証する。この際、外部評価者による導入後の探索的評価の実施に関する AI 事務局のガイダンスがある場合は、それを考慮する。クローズドソースモデルのプロバイダである署名者には、外部評価者のためのいわゆるセーフハーバー体制 を構築することが含まれる。セーフハーバー体制には、独立したモデル評価のために、展開モデルへの安全かつ非 制限的なアクセスを容易にすることが含まれるが、これには、確立された関与ルールと悪用を防ぐセーフガードの

遵守 が条件となる。このようなアクセスプロトコルは、強固なモデル評価の必要性と、モデルおよび評価環境の完全性と機密性を維持する必要性とのバランスをとるものでなければならない。無償のオープンソースライセンスの下でモデルをリリースすることも、このようなアクセスを提供する方法の一つである。オープンソースではなく、オープンウェイトでモデルを公開する場合、署名者はそのモデルで行われる研 究を促進することを約束する。

下位施策 16.2 には、責任ある情報開示の枠組みに沿うものであれば、評価者や評価結果の公表に制限を課 すことなく、専門家の独立性を支援することが含まれる。外部評価者が、識別されたリスク、テスト方法、緩和案の概要を記した報告書を署名者に提出する場合、 署名者はそれをモデル報告書に含めることを約束する。さらに署名者は、そのような証拠を用いて、コミットメント 6.4 と 10.2 を満たすことが奨励される。

署名者は、バグ報奨金プログラムなど、市場導入後の外部アセスメントを積極的に奨励する。

主要業績評価指標の可能性

- **KPI 16.2.1：署名者は、外部評価者のための安全かつ非制限的なアクセスプロトコルを文書化し、実施/公表する。 これには、関与のルール、責任ある情報開示の原則、報告要件が含まれる。**
- **KPI 16.2.2：署名者は、明確な成功規準、金銭的または計算ベースの報奨金、および適切な非報復の保証を伴うバグ報奨プログラムを確立している。**
- **KPI 16.2.3：KPI16.2.3：施策 16.2 に基づく独立した外部評価者は、例外的な状況を除き、正式な要請から 30 日以内に展開モデルへのアクセスを受ける。**
- **KPI 16.2.4：署名者は、調査結果が責任ある情報開示の枠組みを遵守していることを条件に、外部評価者が 60 日の通知期間後に調査結果を公表することを認める方針を文書化している。**
- **KPI 16.2.5：署名者は、すべての独立した試験結果をモデル報告書に記載している。**
- **KPI 16.2.6：署名者は、外部評価者が誠意をもって行動し、業務規程を遵守している場合に、 法的または金銭的な影響から保護するためのセーフハーバーポリシーを採用している。**

施策 16.2 の変更点の説明：初稿では、この対策に反対する回答者はほとんどいなかった。そのため、今回のドラフトでは、セーフハーバーとはどのようなものかを概説することで、この施策をさらに明確にすることを目指した。

コミットメント 17. 重大インシデントの報告

法律文書

AI 法第 55 条第 1 項（c） :「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、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。

AI 法第 55 条 1 項(c)の義務を果たすため、署名者は、適切なリソースを確保したプロセスを設置する。

署名者は、インシデントの重大性、インシデントがどの程度進行中であるか、AI 事務局やその他の外部関係者が対応可能かどうかに照らし、何が不当な遅延となるかを理解する。現在進行中のインシデントについては、特に報告の遅れを避けることを約束す

る。

他の AI 事務局ガイダンスがない場合、重大インシデントは AI 法第 3 条 49 項に従って定義される。例えば、人の死亡、人の健康への深刻な危害（AI 法第 3 条 49 項(a)）、重要インフラの管理または運用の深刻かつ不可逆的な混乱（AI 法第 3 条 49 項(b)）、財産または環境への深刻な危害（AI 法第 3 条 49 項(d)）などである。

法律文書

AI 法第 3 条第 49 項重大インシデント」とは、直接的または間接的に以下のいずれかにつながる AI システムのインシデントまたは誤作動を意味する：

(a) 人の死亡、または人の健康への重大な危害

(b) 重要インフラの管理・運営に深刻かつ不可逆的な混乱が生じること。

(c) 基本的権利の保護を目的とする EU 法の義務を侵害するものである

(d) 財産や環境に重大な損害を与える。

このコミットメントには、重大なインシデントへの直接的または間接的なモデルの関与が疑われる場合、それを調査するために十分なリソースを割くことが含まれる。さらに、疑惑が確認された場合、 署名者はそのインシデントを AI 事務局に遅滞なく報告することを約束する。署名者が AI 事務局にインシデントを報告しても、不正行為を認めることにはならない。同様に、署名者は、重大インシデント報告書を最終文書として扱わない。

署名者は、ビジネスモデルと展開手段を考慮した上で、適切な方法を用いて重大なインシデントを特定する。例えば、以下のような方法がある：

- サードパーティによる重大インシデントの報告を、AI 事務局及び／又は各国所轄官庁に直接、又は署名者を通じて間接的に促進する。
- AI 事務局および／または各国の所轄当局に直接、または署名者を通じて間接的に、API 顧客など川下のプロバイダやユーザーからの重大インシデント報告を促進する。
- 該当する場合、問い合わせや世代に関するユーザーログをアセスメントする。
- 電子透かしや、出力をモデルまで遡るためのその他の技術を使用して、モデルによって生成されたコンテンツを識別する。
- 警察やメディアの報告書、ソーシャルメディアへの投稿、ダークウェブ・フォーラム、外部機関による調査など、インシデントの証拠となる他の情報源を検討する。

重大なインシデントの特定に優先順位をつけ、より重大度が高く、モデルが間接的または直接的に関与している確証 が高いインシデントにより多くのリソースを費やす。署名者は、重大インシデントの特定、報告、対応の過程で作成された文書を、少なくとも 12 カ月間保存する。

署名者は、AI 事務局が提供する重大インシデント報告に関するガイダンスやテンプレートが利用可能な場合、それ に従うことを確約する。それらがない場合、署名者は以下の要素を報告する：

- インシデントの開始日と終了日、またはその概算。

- インシデントの性質（結果として生じた損害を含む）。
- インシデントについて、それに至る一連の出来事を含めて説明すること。
- 可能な限り、モデルのアウトプットとその生成に至った要因の説明を含む根本原因の分析。これには、使用されたインプットと、これらのアウトプットを生み出す一因となったセーフガードの潜在的な失敗や回避を特定することが含まれる。
- 直接的であれ間接的であれ、インシデントに関与した可能性のあるモデルである。
- 重大なインシデントへのモデルの直接的または間接的な関与に関して入手可能な証拠の説明。
- これに対し、署名者は何をするつもりなのか、または何をしたのか。
- 署名者が、AI 事務局または各国所轄官庁に、これに対して何をするよう勧告するのか。

署名者がまだ特定の関連情報を持っていない場合、重大インシデント報告書にその旨を記載する。インシデントがすでに NIS2 など他の枠組みを通じて報告されている場合、 署名者はその報告書を代わりに送付することができる。

署名者は、AI 事務局からのガイダンスがあれば、それに沿って、重大インシデントが発生しそうになったニアミスも報告することが奨励される。

コミットメント 17 の変更点の説明：対応者はコミットメントに好意的であったが、何をもって重大インシデントとするのか、どのように特定するのか、どのように報告するのかを明確にするよう求めた。これらの点を明確にするよう努めた。

さらに、最初のドラフトでは、コミットメント 17 と施策 17.1 の間に食い違いがあった。現在、ニアミスの報告は「奨励される」行動として追加されている。

初動対応者の中には、「重大インシデント対応準備」は規程の他の部分により適合するため、初動ドラフトの施策 17.2 を削除した方がよいとの意見もあった。これを受けて、本ドラフトでは、この施策を施策 4.5 に移した。

コミットメント 18. 内部告発の保護

法律文書
第 87 条 AI 法"指令（EU）2019/1937 は、本規則の違反の報告及び当該違反の報告者の保護に適用される。"

AI 法第 87 条に基づく義務を果たすため、従業員 50 人以上の署名者は、指令（EU）2019/1937 に基づき、内部告発手段を導入し、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。

署名者は、指令（EU）2019/1937 に続く内部告発保護に関する義務を認識する。この認識は、これらの義務の完全な適用と国内法への移行の方法を妨げるものではない。以下の規定は、最も重要な義務を強調することを意図したものであり、指令（EU）2019/1937 に基づくすべての責務の網羅的なリストを構成するものではない：

- 書面と口頭の両方で報告できる、安全で機密性の高い内部報告手段を確立する。
- 公平な担当者または部署を指定し、報告を処理し、7 日以内に受領を確認し、3 カ月以内にフィードバックを提供する。

- 機密性を確保し、GDPR に基づく仕様を含むがこれに限定されないデータ保護要件を尊重しながら、適切な記録を維持すること。
- 内部および外部の報告手続きに関する明確な情報を提供する。
- 解雇、降格、嫌がらせなど、あらゆる形態の報復から内部告発者を保護する。

AI 事務局が内部告発のための指定メールボックスを設置した場合、署名者はその存在を対象者に通知する。

従業員数が 50 人未満の署名者は、上記の内部告発者保護を実施することが奨励される。特に、AI 事務局のメールボックスが運用されている場合は、少なくとも年 1 回従業員に通知し、解雇、降格、嫌がらせを含むあらゆる形態の報復から内部告発者を保護するための対策と方針を実施する。

主要業績評価指標の可能性

- **KPI 18.1：従業員が、報復を恐れることなく、自由に、匿名で不正行為を報告できることを保証する内部告発方針の証拠を文書化する。この方針には以下が含まれる：**
- **保護される開示の範囲。**
- **内部および外部の報告チャネル。**
- **報復防止施策と保証。**
- **報告された懸念事項の調査プロセス。**
- **司法権の余裕と保護の限界。**
- **KPI 18.2：保護 18.2：内部通報の保護および通報経路について、従業員およびステークホルダーへのコミュニケーション戦略の記述。**
- **KPI 18.3：少なくとも年 1 回実施される、定期的な意識向上活動（研修、社内コミュニケーション、キャンペーンなど）の証拠。**
- **KPI 18.4：システミック・リスクを伴う汎用 AI モデルに関連する懸念に特化した、匿名の内部報告メカニズム（コンプライアンス・ホットライン、安全なメッセージング・プラットフォームなど）の提供。**
- **KPI 18.5：機密性を保持しつつ、懸念の種類ごとに分類された、AI 関連の内部通報を文書化する。**
- **KPI 18.6：初動対応時間と報告された懸念事項への対応に要した時間を文書化する。内部通報の初動対応と確認は過度な遅延なく行われ、内部通報に対する施策は通報された懸念の重大性に比例する。**

コミットメント 18 の変更点の説明：様々な利害関係者から、指令（EU）2019/1937 に基づく最も重要な義務を、この規範の中で繰り返し説明すべきであるという意見が出された。さらに、同指令に該当しない小規模企業に対し、調整された最小限の内部通報保護を実施することを奨励する文言が追加された。

コミットメント 19. 通知

AI 法第 52 条 1 項および第 55 条 1 項に基づく義務を果たすため、署名者は、システミ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、その枠組み、モデル報告書、適切な場合には実質的なシステミ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、AI 法第 78 条に従って提供された情報の秘密を保護する AI 事務局の義務を理解した上で行われる。

コミットメント *19* を達成するために、以下を行う：

施策 19.1 システミック・リスクを通知する汎用 AI モデル

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ることを約束する。その計算能力が AI 法第 51 条 2 項に規定される閾値を超えた場合、署名者は 2 週間以内に AI 事務局に通知する。署名者は、分類基準を満たすモデルを EU 市場に投入しないと確信する強い理由がある場合、AI 局に通知しないこともできる。

計算力の見積もりは、ベストプラクティスや AI オフィスからのガイダンスがある場合はそれに従って行われる。

システミック・リスクを伴う汎用モデルに分類されるモデルを上市することを AI 事務局に通知する際、 署名者は、現行の枠組みが遵守された場合、当該モデルに対して適切であると考えるかどうかも明記する。

下位施策 19.1 の変更点の説明：下位施策 19.1 は、主に、AI 法第 51 条(1)の(a)の条件が満たされることを署名者がいつ知ることになるかを明確にするものである。これによって、プロバイダがモデルの訓練に使用する計算量を見積もる義務が追加されるのではないかという懸念があった。しかし、我々の理解では、このような訓練の実施には多大なコストとリソースがかかるため、企業が必要な訓練の妥当な見積もりなしに訓練を実施することは考えにくい。今回の変更は、この理解を反映したものである。さらに、EU 上市を目的とせず、内部目的または予備実験としてのみ訓練されるモデルもあることを、今回の変更で認めている。

プロバイダはまた、モデル訓練中の計算使用量を見積もるための明確な標準がないことについても懸念を表明した。そこで我々は、標準が存在する場合には、標準に従ってそのような推定を行うべきであることを明確にした。

51(3)にあるように、システミック・リスクを有する汎用 AI モデルの分類規準に国内官庁が加える可能性のある変更を考慮し、本施策を更新する必要があるかもしれない。

施策 19.2. 枠組み更新の通知

署名者は、更新が確認されてから 5 営業日以内に、タイムリーな方法で、AI 事務局がその枠組みの修正されていない最新版を入手できるようにすることを約束する。更新確認プロセスは、枠組みに記載される。このようなアクセスは、一般にアクセス可能なリンクを通じて、または AI 事務局が指定する十分に安全な経路を通じてプロバイダが提供することができる。

施策 19.3. 枠組みの妥当性評価通知

署名者は、枠組み妥当性評価の最新版（修正なし）を、AI 事務局が適時に、遅くとも評価が確定してから 5 営業日以内に入手できるようにすることを約束する。

施策 19.4. 安全・安心モデルレポート通知

署名者は、システミック・リスクを伴う汎用 AI モデルを上市するまでに、AI 事務局が指定する十分に安全なルートを通じて、安

全・安心モデル報告書を AI 事務局に送付することを約束する。署名者は、システミック・リスクを伴う汎用 AI モデルを市場に投入する 2 週間前までに、安全・安心モデル報告書を共有することが奨励される。

第 19.4 項の変更についての説明：プロバイダは、AI 事務局に安全・安心モデル報告書を提出することが、展開前に承認を必要とすることを示唆しているのではないかと懸念していた。これに対処するため、モデル報告書の提出期限は市場投入前ではなく、市場投入時であることを明確にした。

コミットメント 20. 文書化

第 53 条(1)項(a)および AI 法第 55 条(1)項に基づく義務を履行するため、署名者は、モデルライフサイクルを通じて、 規程遵守に関連する情報を文書化し、要請に応じて AI 事務局および各国所轄官庁に提供する。利用可能な場合、署名者は、AI 事務局が提供する報告用テンプレートまたは API を利用し、また標準化、 政府、業界団体と協力し、異なるプロバイダからの報告結果をできるだけ容易に比較できるよう、また適宜、 含まれるデータを自動分析できるよう、そのようなテンプレートを開発し、それに従うことを約束する。署名者は、AI 事務局からの文書提出の要請に迅速に対応し、要請があれば、さらなる文書提出や面談を通じて説明を行うことを約束する。

施策 20.1. AI 法第 51 条に基づく分類に関する文書

署名者は、AI 法第 51 条に基づき、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情報を文書化することを約束する。

具体的には、次のとおり：

- コミットメント 1 に概説されている文書要件。
- コミットメント 6～10 に記載されているリスクアセスメントの結果、モデルの評価結果、モデルの一般性、自律性、ツールや足場を使用しアクセスする能力の評価（附属書 XI 第 2 項 1.および附属書 XIII AI 法を参照）。これは、モデル報告書に記載することができる。
- モデルの範囲、特に当組合で設立された登録ビジネス・ユーザーの数（附属書 XIII、ポイント(f) AI 法）。

施策 20.2. 実務規範および AI 法の遵守に関する文書化

署名者は、規範の遵守に関連する情報を文書化することを約束する。これには以下が含まれる：

- 彼らの枠組みだ。
- 彼らのモデルレポート
- コミットメント 15 に記載されているアセスメントとアドヒアランス。
- 該当する場合、ソフトウェアコンポーネントがどのように構築され、または互いに連携し、全体的な処理に統合されるかを説明するシステムアーキテクチャの詳細な記述（附属書 XI、セクション 2、ポイント 3。 AI Act）。これには、入出力フィルタ、微調整、システムプロンプト、機能の抑制／活性化、クエリルーティング、隠れた思考の連鎖、ツールの使用などが含まれる。

- 内部通報者保護方針、および内部通報者の保護と通報経路について従業員や利害関係者に周知していることの保護。

コミットメント 20 の変更点についての説明：このコミットメントは、コミットメント 1 でさらに詳述した AI 法第 53 条に基づき、汎用 AI モデルのプロバイダがすでに直面している義務との一貫性を確保し、重複を避けるために明確化された。文書化要件が複雑であるため、ところどころダブルカウントしていたり、見落としている可能性がある。

コミットメント 21. 公共の性

署名者は、特にシステミックリスクのアセスメントと緩和に関する科学が発展途上であることを踏まえ、一般市民や 外部研究者を含む広範なエコシステムが、汎用 AI モデルから生じるシステミックリスクをよりよく理解し、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。少なくとも署名者は、その枠組み、モデル報告書、または同様の文書（モデルカードやシステムカードなど）を公表することを約束する。安全緩和の有効性を損なうような情報の開示や、社会的利益に不釣り合いな商業上の機密情報の漏洩など、システミック・リスクを実質的に増大させるような情報を含める場合は、情報を冗長化することができる。

主要業績評価指標の可能性

- **KPI 21.1：署名者は、枠組みを AI 事務局に送付してから 15 営業日以内に、枠組みの更新版または新版を公表する。その際、情報を含めることでシステミックリスクが大幅に高まる場合や、社会的利益に不釣り合いな程度に商業上の機密情報が漏洩する場合は、情報を修正する可能性がある。**
- **KPI21.2：署名者は、関連モデルが EU 市場で公表されてから 5 日以内にモデル報告書を公表する。その際、情報を含めることでシステミックリスクが大幅に高まったり、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れたりする場合は、情報を修正する。**
- **KPI 21.3：署名者は、関連するモデルが EU 市場に公表されてから 5 日以内に、モデルレポートの内容を要約した 文書を公表する。例えば、システミックリスクアセスメントの結果や緩和策を記載したモデルカードやシステムカードを公表する。**

コミットメント 21 の変更点についての説明：このコミットメントは、モデル報告書の公表は過度な負担であり、AI 法の要件を超えているとのプロバイダからのフィードバックを受けて改訂された。この懸念に対処するため、社会的利益が最小限の場合には、機密性の高い商業情報の編集を認め、開示要件を緩和した。また、モデル・カードやシステム・カードは、システミック・リスクを伴う汎用の AI モデルとともに公表されることが一般的なモデル・レポートの要約版や冗長版の代わりとなりうることを明確にした。