汎用 AI 実務規範第 3 ドラフト

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

安全・安心セクション

安全・安心部会長および副部会長による序文

私たちは「安全・安心セクション」に関する説明文書と FAQ を作成し、読者はこのウェブページからアクセスできるようにした。以前のドラフトに関わったことがあるかどうかにかかわらず、すべての読者にこれらの文書を読むことを勧める。

実践規範の「安全・セキュリティ」セクションには、主要な AI 企業が AI 法に準拠するための一つの方法が記載されている。AI 法は、EU が 2024 年に可 bi 決した拘束力のある AI 規制である。AI 法のうち、汎用 AI モデル（GPAI）の義務に関する部分は 2025 年 8 月 2 日に発効する。GPAI プロバイダがこれらの義務を遵守するための 1 つの方法の青写真を示すために、欧州委員会は、産業界、学界、政策立案の経歴を持つ独立した AI 専門家グループの一員として、GPAI 実践規範のドラフトを主導するよう私たちを任命した。実務規範の起草は、企業やその他の利害関係者が何度もドラフトを作成し、その都度意見を交換する反復プロセスである。この「安全・保安」セクションは、第 3 次ドラフトの一部である。この「安全とセキュリティ」セクションは、第 3 回目の、そして最後の草案の一部である。このセクションでは、AI 法の GPAI 規則の 1 つのサブセットである「システミック・リスクを伴う GPAI（GPAISR）」規則への準拠方法を規定している。

すべての関係者の貴重なご意見と協力的な姿勢に感謝したい。我々は、2025 年 3 月 30 日（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。ドラフト作成の次のステップに期待したい。

マティアス・サンヴァルト

第 *2* 作業部会

ヨシュア・ベンジオ

第 *3* 作業部会

マリエ・シャーケ

第 *4* 作業部会

マルタ・ツィオージ

第 *2* 作業部会 副議長

ダニエル・プリヴィテラ

第 *3* 作業部会

アンカ・ロイエル

第 *4* 作業部会 副議長

アレクサンダー・ザッヘル

第 *2* 作業部会 副議長

ニタルシャン・ラジクマール

第 *3* 作業部会 副議長

マルクス・アンデルユング

第 *4* 作業部会 副議長

安全・安心セクションの前文

以下の通り：

(a) GPAISR のプロバイダは、モデルライフサイクル全体を通じて適切な対策を講じ、AI バリューチェーンに おける関係者（GPAISR の影響を受ける可能性のある利害関係者等）と協力し、考慮し、システミックリスクマネジメ ントを最新の対策に基づいて構築し、改善されつつある能力に照らして定期的に更新することにより、将来 に備えておくべきである（AI 法 114 項参照）。システミック・リスクのアセスメントは多段階のプロセスであり、GPAISR のシステミック・リスクのアセスメントに用いられる様々な手法を参照したモデル評価は、全体を通じて不可欠である。システミックリスク緩和策が実施されている場合、署名者は、内部または外部の評価者による人的なレッドチ ーム、自動化されたレッドチーム、および/または安全性緩和策の脆弱性を発見し報告するよう奨励するバグ報奨 プログラムなど、最新のプロセスに従い、対象となるシステミックリスクを十分に緩和するための堅牢性と妥当性を継続的 に評価することの重要性を認識する。

(b) しかし、システミックリスクのアセスメントと緩和には、合理的に予見可能な限り、推論時に利用可能なコンピューティングリソースだけでなく、モデルが統合される可能性のあるシステムアーキテクチャやその他のソフトウェアコンポーネントも含めるべきである。GPAISR の輸入事業者が、GPAISR に基づく AI システムも開発、提供又は展開する場 合、GPAISR は、システミックリスクのアセスメント及び緩和において、これらの AI システムを考慮することの重要性を認識する。

(c) 署名者は、システミックリスクのアセスメントと緩和策、および署名者が本規程の規定を遵守していること、または AI 法に基づく義務を遵守していることを文書化し報告する際の精査と詳細の程度は、モデルのライフサイクル全体の関連時点におけるシステミックリスク（AI 法第 56 条(2)(d)）に見合ったものでなければならないことを認識する、例えば、システミック・リスクのレベルが高いほど、モデルの能力、傾向及び／又は効果の範囲に 関する不確実性が大きいほど、及び／又はプロバイダの対応する専門知識が不足しているほ ど、システミック・リスクのアセスメント及び緩和施策は、より徹底的かつ包括的であるべきである。逆に、GPAISR が、システミックリスクが顕在化することなく、既に市場で安全に入手 可能な GPAISR と同じ能力、傾向、または効果を示すと客観的かつ合理的に想定でき、適切な 緩和策が実施されている場合には、より徹底的かつ包括的な対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差異を考慮し、 比例の原則を認識するため、新興企業を含む中小企業にとって、適切な場合には簡易な準拠方法が可能であるべきである（AI 法第 56 条(5)）。

(d) 多くのシステミックリスクアセスメント手法には多大な作業負荷とコストがかかること、またシステミック リスクは複数の AI モデルや AI システム間の相互作用の影響によって左右される可能性があることを、署名 者は認識している。特に、署名者の GPAISR のシステミック・リスクを効果的に評価・緩和するために必要な場合、モデル評 価、ベストプラクティス、インフラを共有すること、あるいは、適切な場合には、独立した外部アセスメ ントと協働することで、「負荷を共有する」ことの利点を認識する。

(e) 中小企業である署名者が、本「安全・セキュリティ」セクションの特定の約束や施策から明示的に除外されている場 合でも、そのような署名者は、自主的にそれらを順守する可能性があることを認識する。

(f) 署名者は、すべてのコミットメントや対策が、当規範の目的、特に目的 IV に照らして解釈される必要があることを認識する。また、当規範の「安全・セキュリティ」セクションの「約束」または「施策」に登場する用語のうち、「安全・セキュリティ用語集」で定義されているものは、同用語集に規定されている意味を持つものとする。

(g) システミック・リスク分類法は、疑義が生じた場合、(1)AI 法第 3 条第 2 項に定義される各リスクの重大性と蓋然性、(2)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、誠実に解釈されるべきである。

(h) 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システミックリスクについて、予防原則（TFEU191 条）の重要な役割を認識し、システミックリスクの特定には、GPAISR の現在の採用率や研究開発の軌跡の外挿を考慮する。

説明欄この安全・安心セクションの前文

このドラフト 3 には、以下の内容が含まれている：

(1) 本安全保証部に関する上記の前文リスト。

(2) 以下の一部の施策および附属書 1 のシステミック・リスク分類法の本文中にイタリック体で示し、太字の表題を付した。

第 2 作業部会 から 第 4 作業部会の議長は、その考え方とアプローチをコミュニケーションしやすくするた め、このドラフト 3 では、イタリック体で書かれた箇所を上記のリストに統合するのではなく、意図的に現在の位置に残している。次回の規程ドラフトでは、すべての前文を上記のリストに統合し、必要に応じて修正する予定である。

GPAISR プロバイダによる一般的責務

コミットメント II.1. 安全とセキュリティの枠組み

法律文書：AI 法第 55 条第 1 項および第 56 条第 2 項（d

署名者は、安全・セキュリティの枠組み（以下、「枠組み」）を採用し、実施することを約束する：(1) 署名者の GPAISR に適用する。(2) 署名者の GPAISR に起因するシステミック・リスクを許容範囲内に抑えるため、システミック・リスクのアセスメント、システミック・リスクの緩和、ガバナンス・リスクの緩和策と手続きを詳述する。

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、ハイレベルのテンプレートを提供する。

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. システミック・リスクの受入規準、システミック・リスク階層、予測 施策 II.1.2-II.1.3

2. システミック・リスクのアセスメントと続行の決定

コミットメント II.2-II.5

3. 技術的なシステミック・リスク緩和

コミットメント II.6-II.7

4. ガバナンスリスクの緩和

コミットメント II.8-II.16

コミットメント *II.1.*を達成するために、以下を行う：

施策 II.1.1. 枠組みの内容

署名者は、約束 II.1-II.16 を遵守するために使用、採用、実施される施策や手順を枠組みに文書化する。

施策 II.1.2. システミックリスクの規準

枠組みにおいて、署名者は、GPAISR に起因するシステミック・リスクが許容されるか否かを判断する規準（「システミック・リスク許容規準」）を記述し、正当化しなければならない。システミックリスク許容基準は、以下のとおりとする：

(1) は、システミック・リスク識別の一環として特定されたシステミック・リスクのそれぞれについて定義される（コミットメント II.3 による）

(2) 少なくとも附属書 1.1 で選択されたシステミック・リスクの種類ごとに、以下のシステミック・リスク階層を含む：

(a) は測定可能である

(b) は、モデル能力、モデル傾向、有害な結果、有害なシナリオ、予想される危害、リスクの定量的見積もり、またはそれらの組み合わせ（例えば、特定された緩和策の記述との組み合わせ）の観点から定義され、かつ、特に適切な安全・セキュリティ上の緩和策がない場合（コミットメント II.6-II.7 による）、その種のシステミック・リスクが容認できないと考えられるシステミック・リスク階層（「容認できないシステミック・リスク階層」）を少なくとも一つ含む；

(3) 上記（2)(c)に従い、許容できないシステミックリスクの階層とは別に、モデルに起因するシステミックリスクを許容できないと判断する方法を記述する。また、関連する広く認知された国際機構、団体、専門家グループによって承認されたベストプラクティス、または利用可能な場合は AI 事務局のガイダンスに合わせる。

さらに、すべてのシステミック・リスクの階層について、枠組み内で、システミック・リスクの階層が有害な結 果とどのように関連しているか、例えばシステミック・リスクのシナリオを説明するなどして、その階層をどのように選択したかについて、署名者は詳細な根拠を示すものとする。

また枠組みにおいて、署名者は以下のことを行う：

(1) 関連するシステミック・リスク階層に関連するシステミック・リスクを軽減することを意図した技術的なシステミック・リスク緩和（コミットメントⅡ.6～Ⅱ.7 に基づき必要なもの）を、科学の現状を考慮した上で可能な限り詳細に記述すること（例えば、GPAISR がシステミック・リスク階層に達した時点で緩和を実施することもあれば、GPAISR がシステミック・リスク階層に達することを防止するために緩和を実施することもある）；

(2) システミック・リスク緩和の現在及び将来の有効性に関する不確実性を含む、経験則上の不確実性を代表する十分な幅の安全マージンを考慮した上で、緩和がシステミック・リスクを許容可能なレベルまで低減するかどうかを署名者がどのように評価するかを記載する

(3) 緩和策が失敗することが合理的に予測できる条件など、緩和策の合理的に予測できる限界を記述すること；

(4) 適切な場合には、所定のシステミック・リスク階層について、1 つ以上のシステミック・リスクに対する緩和策がまだ存在しない旨を記載

(5) GPAISR の更なる開発、最初の又は継続的な市場での入手可能性、及び／又は使用に関する 条件及び意思決定手順を特定し、記載すること。

(6) システミック・リスクを許容できないレベル以下に抑制するための緩和が不十分であるため、GPAISR の開発を進めない、（一般に公開されていない場合には）市場で入手可能とする、及び／又は GPAISR の使用を中止し、モデル評価、市場導入後のモニタリング、又はその他のシステミック・リ スクアセスメント情報に照らして必要な場合には、他のモデルに回帰するための施策を詳述する。

説明欄：システミック・リスクの階層構造

以下は、システミック・リスクの階層に関する 2 つの部分的な例である。これらの例は、2 つの異なる種類のシステミック・リスク階層の構造を説明することを意図している。これらの例は、その内容を保証するものでも、適切な詳細レベルを示すものでもない。

例 1：緩和策がない場合、システミックリスクの階層に到達すべきではない：サイバー犯罪

許容できないシステミック・リスク層：GPAISR は、十分なリソースを持つ国家レベルのサイバー組織と同規模のサイバー攻撃を、人間のサイバー専門家が行うのと同等かそれ以下のコストで完全に自動化することができる。

許容できないリスクレベルに到達する前に必要なセキュリティ緩和：ランド計画 SL4 のセキュリティ目標を満たすセキュリティ緩和。しかし、そのような緩和策はまだ存在せず、活発な研究の対象である。

例 2：システミック・リスク階層に到達すべきではない

システミック・リスク：コントロールの喪失

許容できないシステミックリスクの階層：GPAISR は、学習データにポイズニングを行い、推論規範にバックドアを挿入することで、将来のモデルの開発と利用を妨害する。

リスク階層以下にとどまるための緩和：モニタリング、AI 制御プロトコル、微調整、セキュリティ対策など、モデルが許容できないリスクレベルに達するのを防ぐ緩和策を適用する。

施策 II.1.3. 予測

枠組みにおいて、（II.1.2 項に従って特定される）システミックリスクの階層ごとに、特定のモデル能力に依存する場合、 署名者は以下のことを行うものとする：

(1) 適切なシステミック・リスク緩和の準備を促進するため、当該機能を有する GPAISR を最初に開発すると合理的に予測される時期の見積りを、最善の努力と最新の手法を用い、当該機能が既に市場で入手可能な 署名者のモデルのいずれにもまだ備わっていない場合、記載する

(2) (a)見積もり時期の基礎となる前提条件、(b)見積もり時期の正当化理由、(c)当該見積もりに関する不確実性とその理由。

将来、前文に記載される可能性のある資料：第 *II.1.3* 項を実施するため、署名者は、適切かつ利用可能な場合、業界団体の協力を得て、他の 署名者と協力または組み合わせて、総計予測、調査、その他の推定を利用することができる。

指標 II.1.4. 意思決定における外部意見の透明性

枠組みにおいて、システミックリスクのアセスメントと緩和に関する意思決定に透明性を持たせるため、 署名者は、GPAISR のさらなる開発、市場での最初の入手可能性、継続的な入手可能性、および/または使用が、 関連する政府関係者を含む外部関係者からの情報に基づいて行われるかどうか、また、いつ行われるかを特定し、説明するものとする。

将来、前文に記載される可能性がある：施策 *II.1.4* を満たすためには、外部の関係者からの情報（または認可）は必要ないが、フロンティア *AI* の安全性コミットメント（成果 *3.VIII*）などの既存の国際的なコミットメントに従って、その存在（そしてデフ

ォルトではその不在）にまつわる透明性を確保する必要がある。

施策 II.1.5. 枠組みの改善と更新

署名者は、自国の GPAISR に起因するシステミック・リスクを評価し緩和するための枠組みの有効性を、以下 を含め、長期的に改善する：

(1) 枠組みを自国の GPAISR に適用することで得られた洞察に基づいている

(2) また、枠組みが、システミック・リスクの評価及び緩和のための最もよく知られた手段及び手順を取り入れることを含め、他の枠組み又は公に利用可能な類似のシステミック・リスクの評価及び緩和の方針を考慮に入れ、最先端の技術を採用し、これを考慮に入れることを確保すること。

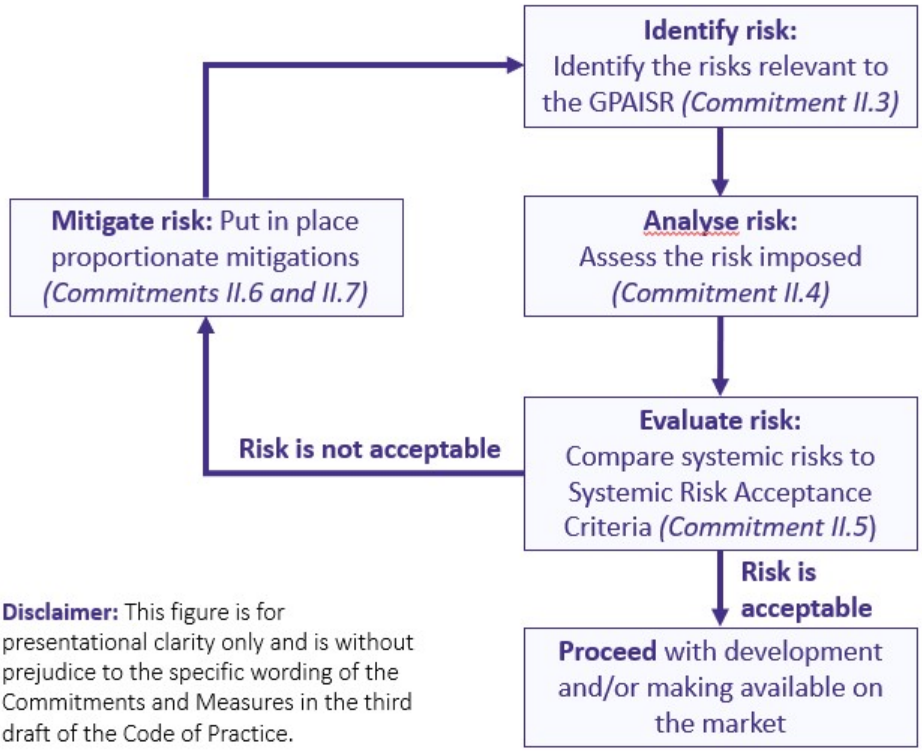
枠組みにおいて、署名者は以下のプロセスを説明する：(1) フレームワークを更新する (2) フレームワークの更新が完了したと判断する。フレームワークが更新された場合、署名者は、変更履歴にバージョン番号と変更日を記載し、フレームワークの更 新方法と理由を説明する。

将来、前文に記載される可能性がある：ただし、枠組みが問題となるシステミック・リスクに相応に対応するものであることを条件とする（本規程の安全・セキュリティ編の前文（c）に従う）。

第 2 ドラフトからの変更点の概要：第 II.1.2 章では、システミック・リスクの階層化を、選択されたシステミック・リスク（附属書 1.1）のみに義務付け、その他のシステミック・リスクについては、ステークホルダーからのフィードバックに示されたとおり、適切でない場合は任意とした。また、システミック・リスクの階層を「署名者間で固定され、比較可能な規模」で定義するという要件は、システミック・リスクの階層をさまざまな方法で定義するという、本施策における署名者に与えられた柔軟性と矛盾するため、削除した。

また、システミック・リスクの階層を予測することから、システミック・リスクの階層に関連する危険な 能力を予測することに要件を絞り込み、業界グループを使って署名者全体で集計できるよう柔軟性を持たせた。これは、結果（広範な経済的損害など）は能力よりも予測しにくいという利害関係者のフィードバックに応えるものである。

How should Signatories assess and mitigate the systemic risk of their general-purpose AI model with systemic risk (GPAISR)?



Disclaimer: This figure is for presentational clarity only and is without prejudice to the specific wording of the Commitments and Measures in the third draft of the Code of Practice.

GPAISR プロバイダのリスクアセスメント

コミットメント II.2. モデル開発時を含め、モデルライフサイクル全体を通じてのシステミックリスクアセスメントと緩和

法律文書：AI 法第 55 条 1 項、前文 110、前文 114

署名者は、システミック・リスクアセスメントを適切な時点で体系的に実施することを約束する。

モデルのライフサイクル全体、特にモデル を市場で入手可能にする前の段階においてである。具体的には、署名者は、GPAISR の開発期間中、「このコミットメントのための施策」に規定されるとおり、システミックリスクの評価と緩和を開始することを約束する。

コミットメント *II.2* を達成するために、以下を行う：

施策 II.2.1. 開発計画

GPAISR に分類される汎用 AI モデル、または署名者が GPAISR の分類条件（AI 法第 51 条(1)(a)に基づく）を満たすことを知り、もしくは合理的に予測できるモデルの開発を計画する場合（訓練を含む）、遅くとも AI 事務局に通知（AI 法第 52 条(1)に基づく）してから 4 週間以内に、署名者は開発を計画するものとする：

(1) コミットメント II.1 に従い）適切な枠組みを整備し、計画されたモデルについてその実施を開始する。

施策 II.2.2. 開発中

事前学習、訓練、事後学習(微調整、強化学習、または類似の方法など、署名者自身または署名者の代わ りに行われる場合)を含む開発期間中、署名者は以下のことを行う：

(1) 訓練開始前に定義され、文書化された適切なマイルストーンにおいて、システミックリスクをアセスメントし、必要に応じて緩和する：

  (a) コンピュート・ベースのマイルストーン（例えば、有効コンピュート数が 2 倍から 4 倍増加するごとに）を訓練する；

  (b) 開発プロセスに基づくマイルストーン（例えば、微調整や強化学習の段階、より多くの個人にモデルへのアクセスを許可する前、ネットワーク、インターネット、ハードウェアへのアクセスなど、より多くのアフォーダンスをモデルに許可する前など）。

  (c) メトリックスに基づくマイルストーン（例えば、事前学習損失や評価実績のレベル）

(2) システミックリスクアセスメントをさらに実施するために開発を一時停止することを正当化するようなシステミックリスクの大幅な変化を特定するための適切な手順を導入する。例えば、高度にスケーラブルでリアルタイムな能力向上の特定を可能にする自動ベンチマークを導入することにより、人的または組織的なボトルネックのリスクを低減する；

(3) 上記(1)に従って定義された各マイルストーンにおいて、次のマイルストーンまでに、モ デルがまだ適切な緩和施策を講じていないシステミックリスク階層に到達することが合理的に予見可能で あるかどうかをアセスメントし、それに応じて、モデルがこれらのシステミックリスク階層に到達する前に、 かかる緩和施策が講じられることを確保すること。

(4) モデルが次のマイルストーンを超える前に、システミックリスクアセスメント及び緩和施策、手順、 決定をすべて文書化する。

将来、前文の素材となる可能性がある：署名者は、規程が「署名者のために行われた行為」と述べている場合、以下の行為を指すと認識する：*(1)*正式に任命された代理人、代表者、従業員（雇用の範囲内で行動する）、*(2)*署名者から当該行為を請け負った者、または請け負った 者、*(3)*金銭的報酬の有無を問わず、署名者の利益のために署名者の指示の下で行動する者。のために」という表現は、独立した川下プロバイダを指すことを意図していない。

第 2 ドラフトからの変更点の概要：ステークホルダーからのフィードバックに基づき、他のコミットメントとの重複を大幅に削除したため、このコミットメントは大幅に短くなった。「アセスメント準備（Evaluator readiness）」は施策Ⅱ.4.11、システミックリスクの状況の変化の可能性を評価するためのトリガーは施策Ⅱ.8.7、セキュリティ対策は施策Ⅱ.7.6 で説明する。これは、特にオープン・ウェイト・モデルについて実施するのが困難であり、リソースを最 も有効に活用していないように思われた、廃止後の GPAISR を監視する要件を削除したことを意味する。また、本規定が不必要な作業を生じさせないようにするため、より柔軟なトリガー規 準を採用し、6 ヶ月ごとの固定サイクルでの再評価を義務付けないこととした。GPAISR の通知を考慮するため、第 II.2.1 施策の枠組み準備のトリガーを明確化した。最大の追加点は、継続的な技術開発を考慮し、規範に将来性を持たせるため、II.2.2 項において、訓練中のシステミック・リスク・アセスメントのマイルストーンを定義する別の方法を導入したことである。

コミットメント II.3. システミック・リスクの特定

法律文書：第 55 条 1 項、第 56 条 2 項（c）および（d）、AI 法第 110 条前文

署名者は、自国の GPAISR に由来するシステミック・リスクを選定し、その特徴をさらに明確化する。

このコミットメントの対策に明記されているように、更なる評価と緩和を必要とするほど重大である。

コミットメント *II.3.*を達成するために、以下を行う：

施策 II.3.1. システミック・リスクの選択

署名者は、附属書 1.1 に記載されているシステミック・リスクの種類を選択し、さらなるリスクアセスメントを行い、 必要であれば緩和を図る。

GPAISR がその他のシステミック・リスクをもたらすと合理的に予測できる場合には、適切な検討を行う：

(1) 附属書 1.2 および 1.3、ならびに、附属書 1.4 および市場で入手可能な類似モデルが示すリスクに関する情報、問題となっているリスクに関する最先端の研究、影響を受ける可能性の高い専門家や利害関係者グループの知識や見解など、その他の関連する検討事項、

署名者はさらに、GPAISR の影響力の高い能力に特化したその他のシステミック・リスクを選択し、さらなるアセスメントを行い、必要であれば緩和を図るものとする。

施策Ⅱ.3.2. システミック・リスク・シナリオの決定

下位施策 II.3.1 に従い選択された各システミック・リスクについて、署名者は、システミック・リスク分析に備え、選択されたシステミック・リスクの種類、性質、発生源をよりよく理解し、特徴付けることを目的に、リスクモデリ ングや脅威モデル（適切な場合は、コミットメント II.4 で理解されるモデル評価を使用）など、適切な手法 を用いてシステミック・リスク・シナリオを作成するものとする。各システミック・リスクについて、少なくとも 1 つのシステミック・リスク・シナリオを策定する。システミックリスクシナリオを決定する際、署名者は、GPAISR の合理的に予見可能な使用と誤用（過失、 無謀、故意の別を問わない）を考慮するものとする。各システミックリスクシナリオには、以下を含むものとする：(1) GPAISR の開発、市場での入手可能性、および/または使用により、選択されたシステミック・リ スクが発生する可能性のあるシステミック・リスク経路（「危害経路」）、および（2）少なくとも附属書 1.4 の潜在的システミック・リスク発生源に基づく、選択されたシステミック・リスクの発生源。

第 2 ドラフトからの変更点の概要：システミック・リスクの識別の範囲とプロセスについて、さらに明確化するようステークホルダーから要請があった。第 II.3.1 項において、システミック・リスク識別とシステミック・リスク分類法（附属書 1）との関連、その他のシステミック・リスクを識別するメカニズム、その選択を決定するための関連する検討事項を明記した。我々は、システミック・リスクを識別するためのガイダンスとサポートを署名者に提供する一方で、要求されるプロ セスが実用的で実現可能なものであるよう、慎重にバランスを取った。さらに、選択されたシステミック・リスクをよりよく理解し、システミック・リスクの特定とシステミック・リスク分析との間の流れを改善するため、システミック・リスク・シナリオの決定を導入する II.3.2 項を新たに追加した。方策 II.3.2 では、システミック・リスク・シナリオの決定に必要なステップと範囲を規定し、リスク・モデリングや脅威モデリングなど、 、有用な具体的手法を提案したが、署名者の検討は必須ではない。

コミットメント II.4. システミック・リスク分析

法律文書：第 55 条 1 項、第 56 条 2 項(d)、前文 110 および前文 114 AI 法

システミックリスクアセスメントの一環として、署名者は、システミックリスクの厳格な分析を実施する。

コミットメント II.3 に従い特定されたシステミック・リスクの重大性と可能性を理解するため、システミック・リスク分析を実施する。署名者は、関連する GPAISR に起因するシステミック・リスクに応じ、また本コミットメントの「施策」に明記されている通り、様々な深度・強度のシステミック・リスク分析を実施する。システミック・リスク緩和が実施される場合、署名者はシステミック・リスク分析の一環として、その有効 性と頑健性を検討する。

本コミットメントの「施策」でさらに規定されているように、署名者は、システミックリスク分析に おいて、モデル・アフォーダンス、安全なオリジネーターモデル、市場で入手可能なモデルおよび/または 使用される可能性のあるコンテクストとその影響を考慮し、モデルに依存しない情報や最新のモデル評 価を含む、様々な情報や手法を活用することを約束する。

将来、前文に記載される可能性がある：署名者が実施するシステミック・リスク分析の程度は、評価されるシステミック・リスクと、ある *GPAISR* がどの程度のシステミック・リスクをもたらすかについての不確実性に比例する。署名者は、システミック・リスク分析の優先順位を、提示されたシステミック・リスクに比例させるため、リソースをあまり必要としない方法（文献調査やオープン・ベンチ

マークの実施など）でシステミック・リスク分 析を開始し、より詳細なモデル評価（エージェント・モデルの評価やヒューマン・アップリフト研究など）でフォローアップすることができる。

コミットメント *II.4.*を達成するために、以下を行う：

施策 II.4.1. システミック・リスクの推定

システミック・リスク分析の一環として、署名者は、コミットメント II.3 に従い特定されたシステミック・リスクについて、GPAISR に起因するシステミック・リスクのレベルを、分析されるシステミック・リスクの種類や性質に応じ て、定量的および／または定性的な推定を用いて測定する。その際、署名者は以下のようなシステミック・リスクの推定技法を用いるものとする：

(1) それは厳格で最先端のもので

(2) また、重大インシデントに関するデータ（コミットメント II.12 による）及び市販後モニタリング（少なくとも施策 II.4.14 による）を含む関連情報に基づく。

これらのシステミック・リスクの見積もりは、II.3.2 項で策定したシステミック・リスク・シナリオに基づくものとし、（コミットメント II.5 に従い）システミック・リスクの見積もりを評価するために、（II.1.2 項の）あらかじめ定義されたシステミック・リスク許容規準に照らして比較可能なものとする。

リスクアセスメントの対象となるシステミック・リスクが可能かつ適切なものである場合、署名者は、これらの推定 値を定量的な用語で記載する。少なくとも附属書 1.1 のシステミック・リスクの種類につい ては、システミック・リスクを測定する定量的な指標となるシステミック・リスク指標を提示し、自国の枠組みで要求されているシステミック・リスク階層への進捗を追跡することで、早期警戒の兆候を予測するため、署名者は最善の 努力を払うものとする（II.1.2 項）。署名者は、システミックリスクの階層ごとに、少なくとも 1 つのシステミックリスク指標を定義することを検討するものとする。

システミック・リスクの種類を評価する上で定量的推定が不十分な場合、署名者は定性的推定で補完する。定量的な見積もりだけではシステミック・リスクの評価として不十分な場合、署名者は定性的な見積もり、または定性的な見積もりと定量的な見積もりを併用するものとする。ただし、これらの見積もりは、（コミットメント II.5 に従い）推定されたシステミック・リスクを評価するために、あらかじめ定義されたシステミック・リスクの受入規準（手段 II.1.2 による）と比較できるものとする。

システミック・リスク推計の結果としては、例えば、（1）モデル評価結果および収集したその他の情報の集計、（2）モデルが想定するシステミック・リスク経路の定性的記述、などが考えられる。

(を可能にする（または可能にしない）（例：「我々のモデルは、専門家が新規の脅威ベクトル X を開発することを可能にする」）、および（3）可能性と予想される損害の定量的推定（例：「我々のモデルは、システミックリスク Z により Y ユーロの損失を引き起こす可能性が X%ある」）。

第 2 ドラフトからの変更点の概要：ステークホルダーからのフィードバックに基づき、リスク推定が何を意味するのか、またリスク識別やリスク評価（現在は「システミックリスクの受容決定」）との関係をさらに明確にした。そのため、冗長性を回避し、明確性を高めるため、「方法論」と「リスク推定」に関する従来の施策を 1 つに統合し、システミック・リスクの推定とシステミック・リスクの受容判断との関連付けを容易にするため、システミック・リスク指標の概念を再導入し、システミック・リスクの推定結果がどのようなものかを具体的に例示した。

施策 II.4.2. 安全に導き出されたモデル

システミックリスク分析作業の重複を避けるため、「安全な派生モデル」についてコミットメント II.3 に従って特定されたシステミックリスクを分析する際、「安全なオリジネータモデル」の既存のシステミックリスク分析から得られた情報を考慮するよう、署名者は最善の努力を払うものとする（例えば、安全な派生モデルについて同じモデル評価を再度実施する代わりに、安全なオリジネータモデルについて実施されたモデル評価の結果を使用するなど）：

(1) 安全なオリジネーターモデル：(a)本規程に従ったシステミックリスクマネジメントプロセ スを経ており、システミックリスク許容規準を満たしている。(c)重大な安全性又はセキュリティ上の欠陥がないこと（例えば、未解決の重大インシデ ントにより確認されたもの）、(d)署名者にとって十分な透明性があること（署名者が、安全緩和、アーキテ クチャ、能力、様式、システミックリスクプロファイルなどのモデルの特性について十分な可視性 を有していることを意味する；

(2) 安全に派生したモデルとは、安全な起点モデルから直接派生したモデル（そのような安全な起点モデルから抽出、量子化、微調整、または後訓練したモデル、あるいは安全な起点モデルの安全性またはセキュリティ緩和を追加または改善したモデルなど）である。

(3) 安全な派生モデルが、安全なオリジネーター・モデルと同じシステミック・リスク・プロファイルを持つか、それよりも低いシステミック・リスク・プロファイルを持つと合理的に想定できる場合。

上記(3)の合理的な前提を確立するための規準は以下の通りである：

(a) ただし、施策Ⅱ.4.11 の(1)に従い、十分な資格を有する評価者がモデル評価を実施していることを条件とする

(b) 安全な派生モデルに対し、一般的な能力の向上、システミックリスクに関連するモデルの能力やアフォーダンスの追加や修正、システミックリスクに関連するモデルの傾向の変更、安全性やセキュリティの緩和の削除や弱体化を意図した、または合理的に予見できる変更は行われていない。

第 2 ドラフトからの変更の概要：本施策は、GPAISR のシステミック・リスクアセスメントにおいて、既存モデルの関連情報を考慮するための厳格な範囲を設定するために追加された。我々は、システミックリスクアセスメント作業が異なるモデルバージョン間で無用に重複しないようにしたいが、ここで「類似モデル」を過度に広義に定義すれば、大きなループホールを開く可能性があることも強く認識している。ご意見をお寄せいただきたい。

施策 II.4.3. モデルに依存しない情報

署名者は、GPAISR およびアセスメント対象のシステミックリスクに関連する、モデルから独立した 情報を収集するものとする。

本施策を実施するため、署名者は以下のことを行う：(1) システミックリスクのレベルに応じ、深度や強度を変化させながら情報の検索と収集を行う：

(a) ウェブ検索；

(b) 文献レビュー

(c) 市場分析（例えば、市場で入手可能な他の GPAISR の能力に焦点を当てたもの）、

(d) 訓練データのレビュー、

(e) 過去のインシデントデータ；

(f) 一般的な傾向の予測（例えば、アルゴリズム効率の発展、計算機使用、データ利用可能性、エネルギー使用に関する予測）

(g) 専門家によるインタビューおよび／またはパネル

(h) インタビュー、調査、地域協議、またはその他の参加型調査手法により、GPAISR が、脆弱性集団を含む、当組合に所在する自然人に及ぼす影響などを調査する。

第 2 ドラフトからの変更点の概要：一般的に使用される方法として、ウェブ検索と訓練データのレビューをここに追加し、より多くの例を示した。KPI は仕様ゲームに脆弱性があるため削除した。

施策 II.4.4. 最新モデルの評価

署名者は、システミックリスクと GPAISR の能力、傾向、効果を適切に評価するため、Q&A セット、タスクベースの評価、ベンチマーク、レッドチームやその他の敵対的試験方法、人間高揚研究、モデル生物、シミュレーション、機密物質の代理評価など、適切な方法論を用いて、最先端のモデル評価を実施することを保証する。

署名者は、モデル評価を実施する：

(1) モデルのライフサイクル全体を通じて、システミックリスクごとに最も適切なタイミングで実施する；

(2) 少なくとも 1 つの関連するシステミック・リスクに関する情報を提供し、関連するシステミック・リスク・シナリオ（施策Ⅱ.3.2 による）のシステミック・リスク指標（施策Ⅱ.4.1 による）を提供する。

可能な限り、署名者は、モデル評価の厳密性を損なうことなく、効率性の高いモデル評価手法を使用するものとす る。その際、あるレベルのシステミック・リスクが、単純な（しかし厳密な；第 II.4.5 項による）モデル評価手法に よって決定的に否定される場合（例えば、単純な評価手法では、あるモデルの能力が、合理的な疑義なしに、 あるレベルのシステミック・リスクを顕在化させるには不十分であることが示される場合）、当該システミック・リスクに対す る追加的またはより高度なモデル評価手法は必要ないものとする。

第 2 ドラフトからの変更点の概要：ここでは、KPI と同様に、仕様の賭博に脆弱性を持つ他の重複を削除した。最新技術」の定義が用語集に記載され、業界のコンセンサスと AIO によるガイダンスの両方を考慮していることに留意されたい。ここでは特に、II.4.5 から II.4.8、II.4.11 に記載されているような関連する品質規準を常に考慮し、「単純な」評価方法が、より複雑な評価を実施しないことを正当化するために使用できる場合について拡大した。繰り返しになるが、ここでの目標は、不必要な作業を防ぎ、「本当に重要なこと」に集中できるようにすることであるが、システミック・リスクを増大させることはない。

施策 II.4.5. 厳密なモデル評価

署名者は、GPAISR において、高い科学的・技術的厳密性を有するモデル評価が実施されることを確 保するものとする。システミック・リスク評価のために実施されるモデル評価においては、予備的・探索的な調査を容易にするために必要な場合、署名者のモデル報告書が II.8.3 項(2)に準拠することを条件に、より低いレベルの厳密性が許容される。

本施策を実施するため、署名者は以下を行う：(1) モデル評価の種類に応じて、機械学習、自然科学、社会科学の科学的査読で使用される品質標準と同等かそれ以 上の品質標準を採用し、評価中のモデルの挙動や性能がソフトウェアのバグに大き

く影響されないようにする など、ソフトウェア工学における最先端の手法を採用する。中小企業の署名者が、それぞれの品質基準または実施方法を特定または実施するための専門知識または資 金的資源を社内に持たないなどの理由で、本項で概説される品質基準または実施方法を採用できない場合、その旨を AI 事務局に通知し、本施策を履行するための適切な代替手段を AI 事務局と合意するものとする。

将来の前文に記載される可能性がある：内部妥当性の達成は、システミックリスクの種類や手法によって異なるが、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 変数の制御と妥当性確認のための相関緩和、学習データの混入がないことの証明、学習におけるテス トデータの使用の防止（学習-テスト分割の使用やカナリア文字列の尊重）、異なる条件下や評価環 境でモデル評価を複数回再実行すること（モデル評価の個々の部分を変化させることを含む。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）を変化させることを含め、異なる条件や評価環境のもとで複数回モデル評価を再実行すること、軌道やその他の出力を詳細に検査すること、あるいは、新しい評価を作成し、その完全性を確保するために管理するプロセスを（公開または *AI* 事務局に）開示すること。

将来、前文に記載する可能性がある：外部妥当性の達成は、システミックリスクの種類や手法によって異なるが、例えば、評価プロセ スにおける専門家の適切な統合、適切な能力引き出し（評価尺度Ⅱ*.4.6* 参照）、評価が実施され る技術的な評価環境の条件や現実の状況との乖離を文書化する（評価尺度Ⅱ*.4.7* も参照）、評価環 境の多様性と現実性（評価尺度Ⅱ*.4.8* も参照）、あるいは適切に保留されたテストセッ トの利用（例えば、最高レベルのシステミックリスクの場合）等によって示すことができる。*.7* も参照）、評価環境の多様性と現実性（施策 *II.4.8* も参照）、または適切に保留されたテストセット（例えば、最高レベルのシステミックリスクについては、隔離されたテストセットと独立したサードパーティによる保留を使用する）を利用することである。

将来の前文に記載される可能性がある：再現性は、例えば、ピアレビューの成功（ ）、独立した第三者による複製、適切な量のモデル評価データの *AI* 事務局への安全な公開（常に拡散リスクを考慮する）、モデル評価規範と評価方法、評価環境、計算環境、およびエリシテーション方法の文書化、または一般に利用可能な *API*、技術評価標準、およびツールの使用によって示すことができる。

第 2 ドラフトからの変更点の概要：移植性の基準は、II.4.10 項の共有規定とともに再現性の基準で既にカバーされているため、ここでは削除した。また、KPI は仕様の賭博に脆弱性があるため、削除した。3 つの規準の定義には多くの小さな変更が加えられたが、最終的な規程を読みやすくするために前文になるようフラグを立てた。また、既存の科学的・工学的品質標準を採用する必要性を大幅に拡大した。これは、完璧とは言えないかもしれないが、システミックリスクアセスメントの出発点としては適切である。

施策 II.4.6. モデル抽出

署名者は、GPAISR のすべてのモデル評価（内部評価か外部評価かを問わない）が、評価されるシステミック・リ スクに適切かつ比例する、最新水準のモデル抽出により実施されることを保証する：(1) 評価対象モデルの現在及び合理的に予見可能な能力、傾向、影響の上限を引き出す、(2) エリシテーション不足のリスクを最小化する、(3) モデル評価中のモデル欺瞞のリスクを最小化する、(4) 誤用主体が関連するシステミック・リスク・シナリオで役割を果たす場合、潜在的な誤用主体の現実的なモデル エリシテーション能力に合致させる（例えば、潜在的な誤用主体の中には、モデルを完全に引き出すことができない者も いるかもしれない）。

将来、前文に記載する可能性がある：適切なモデル抽出方法は、システミックリスクの種類や *AI* モデルの種類によって異なる可能性があり、例えば、微調整及び／又はプロンプトエンジニアリング（危険な能力や傾向を引き出すなど）、モデルスクラッチパッド、

モデルアンサンブル、足場作り（ツールの使用を含む）、適切な場合にはグレーボックス又はホワイトボックスへのアクセス（施策 *II.4* も参照）、セーフガードなしのモデルの使用、基本モデル又は有用なモデルのみの使用、適切な訓練及び推論計算予算の使用（施策 *II.4* も参照）を含むことができる。*.11* も参照）、セーフガードなしのモデルの使用、基本モデルまたは有用なモデルのみの使用、適切な訓練及び推論計算バジェットの使用（施策 *II.4.11* も参照）。

加えて、モデル評価において、モデルの危険な能力を向上させたり、危険な傾向を増加させるなど、モ デル抽出方法が GPAISR のリスクプロファイルを増加させる場合、モデルへの不正アクセスや、抽出されたモ デルによる有害な行為によるリスクを防止するため、（コミットメント II.7 に従い）適切な、潜在的に増加しうるセキュリ ティ対策が講じられることを、署名者は保証するものとする。

第 2 ドラフトからの変更点の概要：ここでは主に 4 つの変更を行った。(1) 特に、システミックリスクシナリオに GPAISR を十分に引き出すことができない悪用行為者が含まれる場合には、目下のシステミックリスクに対する比例性が重要であることを強調した。(2) 重複する内容を削除した。特に、計算バジェット及びモデルへのアクセスは、第Ⅱ.4.11 項に記載されている。 (3) 一部の評価技法によるリスクの増大と、適切なセキュリティ対策（ ）の必要性を追加した。(4) 多くの GPAISR プロバイダの規模や組織構造に対して、「費やしたエンジニアリング時間」のアプローチは不適切であり、また十分ではないため、KPI を削除した。

施策 II.4.7. システムの一部としてのモデル

システミックリスクのアセスメントと緩和に必要な場合、署名者は、GPAISR のモデル評価において、 評価されるシステミックリスクに応じ、合理的に予測可能な AI システムへのモデルの統合を考慮する ことを保証するものとする。

GPAISR を、署名者自身が市場で入手可能な AI システムに組み込むことを計画している署名者は、モデル評価において、システミックリスクおよび/またはシステミックリスク源に関連する場合、当該 AI システムが市場で入手可能な、または市場で入手可能な、または市場で入手可能な、または市場で入手可能な AI システムが使用される将来の状況を考慮し、モデル評価時に同種の足場やツールを使用するなどして、モデルを適切に評価することを保証するものとする。

署名者が、(1)GPAISR を使用せず、AI システムに統合せず、(2)そのモデルを市場で入手 可能とし、(3)そのモデルが統合された AI システムに関して、本規程が要求する厳密なモデル評 価を行うための十分な情報を、他の手段で入手できない場合、(4)GPAISR のシステミックリスクを効果的に評価・緩和できな い、署名者は、ライセンシーや川下プロバイダと協力し、署名者の関連リリース戦略やビジネスモデルに応じて、また 必要であれば中小企業の規模や能力を考慮し、必要であれば AI 事務局の支援を得ながら、関連情報を入手し、それ に基づいて行動するために最善の努力を払うものとする。

GPAISR を統合する AI システムが市場で入手可能、稼動、使用される方法や状況に、当初のシステミックリスク評価で説明されたものから重大な変更があった場合、署名者は必要に応じてこれを考慮し、GPAISR に起因するシステミックリスクの一部または全部の再アセスメントを実施し、適宜、関連する変更を考慮し、その結果、施策 II.8.7.に従いモデル報告書を更新するものとする。

第 2 ドラフトからの変更点の概要：ここでの主な変更点は、システミック・リスクに関する必要な情報をどのように収集するかについて、署名者に柔軟な選択肢を与えることを優先し、「ライセンスによる情報収集」を義務化しないことである。

施策 II.4.8. コンテキスト、モダリティ、代表者モデルの評価

当該システミック・リスクのリスクアセスメントと緩和のために必要であり、 関連するシステミック・リスク・シナリオを考慮する場合、 署名者は、モデルが使用されることが予想されるすべての様式を考慮し、モデルの評価が、GPAISR の予想される使用状況 に合致することを保証するものとする。そのために、署名者は、GPAISR に起因するシステミック・リスクを分析し、例えば以下のようなことを 行うものとする：

(1) 例えば、多言語モデルの言語ベースの評価においては、英語だけでなく、欧州の主要な言語及びモデルがサポートするその他の言語をすべてカバーするよう最善の努力を払う一方、システミックリスクの中には、プログラミング能力など、英語で評価するのが最善と思われる能力に基づくものもあることを考慮する。

(2) 例えば、視覚的な入出力の評価、コンピューター制御やロボット工学における動作などである。

想定される使用状況の適切な考慮を促進し、モデル評価のために、またモデル評価中に、関連する 専門知識にアクセスするために、署名者は、可能であり、かつ評価されるシステミックリスクに関連する場 合、市民社会、学界、および／またはその他の関連ステークホルダーの適切な報酬を得た専門家ま たは一般代表者など、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ものとする。関連する利害関係者、例えば GPAISR の直接的な影響を受ける者がいない場合、署名者は、システミック リスクの評価と緩和の一環として、そのような利害関係者の利益を代表する最も適切な代表者を特定し、協 議するものとする。中小企業が本施策を実施するために関連する利害関係者と協力できない場合、AI 事務局に支援を要請す るものとする。

第 2 ドラフトからの変更点の概要：本対応施策では、これらのプロセスにおける専門家および一般代表者に適切な報酬を支払うことを要求している。さらに、文言の明確化を図った。

施策 II.4.9. 探索的調査とオープンエンドのレッドチーム

システミックリスク評価の現状は未発達であり、GPAISR のシステミックリスクを効果的に評価または緩和する ために必要な場合、署名者は、既知のモデル能力、傾向、アフォーダンス、影響に関するシステミックリスク 評価のみに限定せず、GPAISR のシステミックリスク評価または緩和を改善するための（内部および/または外部での） 探索的作業を確実に実施するものとする。

本施策を実施するため、署名者は、組織の規模や問題となっているシステミックリスクに応じて、システミックリスクのアセスメントと緩和手法の最新技術を導入する：

(1) 新たなモデル評価手法の改善、「モデル評価の評価」に関するメタ研究、予測を支援する研究な ど、システミックリスクのアセスメントと緩和に関する探索的研究、市民社会、学術界、その他関 連する利害関係者の専門家または一般代表者を参加者とするオープンエンドのレッドチーム。関連する利害関係者、例えば GPAISR の直接的な影響を受ける者がいない場合、署名者 は、そのような利害関係者の利益を代表する最も適切な代表者を特定し、従事させるものとする。

第 2 ドラフトからの変更点の概要：第 2 ドラフトでは深みが不足していた探索的調査の例を追加した。また、パーセンテージベースの KPI は仕様のゲーム性に脆弱であったため、削除した。

施策 II.4.10. ツールとベストプラクティスの共有

システミックリスクは、インシデントや誤作動の連鎖反応など、他の AI モデルや AI システムとの相互作用に影響されるため、また、署名者の GPAISR のシステミックリスクを効果的にアセスメントまたは緩和するために必要な場合、中小企業でない署名者は、

AI エコシステムにおける関係者の安全・セキュリティ活動を促進するため、最先端のモデル評価（モデル抽出を含む）、システミックリスクのアセスメントと緩和のためのツールやベストプラクティス（確立された方法論や手順など）を共有する、特に、他のモデルプロバイダー（特に中小企業）、川下プロバイダー、独立した外部アセスメント機関、学術機関など、署名者の GPAISR が合理的に相互作用することが予想される AI モデルや AI システムのリスクアセスメントや緩和に携わる者を対象とする。

システミック・リスクのアセスメントに使用されるモデル評価データ（インプットやアウトプットなど）を追加的に共有することを選択した場合、署名者は、この共有を適切に制限するものとし、特にデータを共有してはならない：

(1) 例えば、共有された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場合などである

(2) 科学的厳密性を脅かす可能性がある場合、例えば、共有されたテストデータが訓練セットに漏れる可能性がある場合、あるいは、現在アクティブなテストセットの重要な部分が公開される可能性がある場合、および／または

(3) 公共の安全や商業上の機密情報を保護するために、これらの利益がシステムリスクのアセスメントや緩和の改善による安全上の利益を上回る場合。

非 SME の署名者は、以下を行う：(1) 本施策で言及されるツールやベストプラクティスを共有することにより、署名者自身の GPAISR のシステミックリスク評価と緩和の能力が低下したり制限されたりすることがないようにする。

本規定で言及されるツールやベストプラクティスの共有が、AI 事務局や、AI 事務局が承認する他のイニシアティブ（AI 安全機構国際ネットワークなど）により促進されていない（まだ）場合、署名者の GPAISR のシステミックリスクをアセスメントし緩和するために必要であれば、非 SME の署名者は、可能な限り多くの関係者が関与する方法（ソース規範の公開など）で、または、より限定的な共有が必要な場合は、業界団体が促進できる安全な共有メカニズムを通じて、そのような共有を促進するものとする。

中小企業である署名者は、中小企業でない署名者が共有するツールやベストプラクティスを利用する最善の努力 を払うことで、本項目を満たす。

第 2 ドラフトからの変更点の概要：モデルのシステミックリスクは、他の AI モデルや AI システムとの相互作用によって影響を受ける可能性があることを強調した。システミックリスク評価データの共有は、多くの場合不利であり、システミックリスクの評価と緩和の ために有益な場合にのみ行われるべきである。KPI は、仕様の賭博に脆弱性があり、逆インセンティブを生み出す可能性があったため、削除した。

施策 II.4.11. 有資格のモデル評価チームと、適切な評価へのアクセスとリソース

署名者は、システミックリスクアセスメントを実施するすべてのモデル評価チーム（II.11.1 項に従い、内部およ び/または外部）が、技術的な専門知識と関連分野の知識を組み合わせた学際的なチームであり、評価されるシステミック リスクの全体的な理解を保証する。モデル評価チームは、以下のとおりとする：

(1) 該当するモデル評価作業を実施する資格を有し、それぞれが以下の資格の少なくとも 1 つを満たす複数のチームメンバーで構成される：

(a) 客観的かつ合理的に、システミックリスクアセスメントのためのモデル評 価に貢献する関連専門知識を有するとみなされるような研究又は工学の経 験を有していること（例えば、関連する博士号、関連し認知され査読を受けた論文、又は当該分野への同等の貢献によって実証される）；

(b) アセスメントされたシステミック・リスクに関する、査読を経た、または広く利用されている公表済みのモデル評価を設計または開発した。

(c) アセスメントされたシステミック・リスクに直接関連する分野で少なくとも 3 年間の実務経験があること。

(2) 本規定が求める効果的なモデル評価、GPAISR のシステミック・リスクのアセスメントおよび緩和 を実施するために必要かつ適切な場合には、GPAISR への適切なアクセスが与えられる。当該アクセスは、必要に応じて、モデル能力を引き出すための微調整アクセス、ロジット・ア クセス、十分に高いレート制限、モデルのインプット及びアウトプットへのアクセス（思考 の連鎖による推論を含む）、セーフガードなしのモデルへのアクセス、スクリプトを呼び出した りブラウザを操作したりするモデルの能力といったモデルに対する追加的なアフォーダ ンスをカバーするものとする。さらに、このようなアクセスは、ケースバイケースのアセスメントに基づき、必要に応じ て、グレーボックスアクセス及びホワイトボックスアクセスの技術的実装を対象とする：(a)グレーボックスアクセスは、モデルの「内部構造」を部分的に理解することを可能にする、 ある程度の透明性を提供するものであり、(b)ホワイトボックスアクセスは、モデルの完全な重み、活性化、 その他の技術的詳細が常に閲覧可能であることを意味し、モデルがどのように機能するかについての明 確な洞察を提供するものである。モデル評価チーム（内部および/または外部）にグレイボックスまたはホワイトボックスのモデルへのアクセス権を 与える場合、署名者は、（コミットメント II.7 に従い）このアクセス権（ ）がもたらすモデルセキュリティへの潜在的リ スクを考慮するものとする。署名者は、探索的調査を含め、厳密なモデル評価を実施するのに十分な最低レベルのアクセス権を提供し、より高いレベルのアクセス権を与える場合は、より厳格なセキュリティ緩和策を実施する

(3) 署名者の最善の努力により、評価されるシステミックリスクおよび使用されるモデル評価手法に必要かつ適 切な、モデル仕様、訓練データ、過去のアセスメント結果へのアクセスを提供する

(4) をプロバイダに提供する：(a)十分な長さのモデル評価実行、並列実行、再実行を可能にすることを含む、十 分な計算予算、(b)適切な人員配置、(c)人為的に能力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など、モデル評価結果を精査することを含む、十分 なエンジニアリング予算とサポート、及び客観的に厳密な標準（施策Ⅱ.4.4 参照）に基づ くモデル評価の設計、デバッグ、実行、分析を適切に行うための十分な時間（施策Ⅱ.4.5 参照）が与えられる。.5 を参照）、適切なレベルのモデルエリシテーションを行うことができるようにする（方策 II.4.6 を参照）。与えられた時間は、以下に比例するものとする：(a) 評価されるシステミック・リスクの大きさ、(b) 使用される具体的なモデル評価手法（その手法の新しさ、確立さ れていることの考慮を含む）、(c) 評価対象モデルの安定性及び性能の類似性の程度（市場で入手 可能な最終モデルと比較した場合）：(i)連続するモデルスナップショットのベンチマーク性能が、一般的な能力ベンチマークにおいて有意に変化しなくなった場合、(ii)ツール呼び出しや（半）自律的な意思決定など、モデルの中核機能が変化しなくなった場合。少なくとも 20 営業日というアセスメント期間は、例えば、モデル評価チームがほとんどのシステミックリスクとモデル評価手法に対して十分な時間を与えられたことを示す可能性がある。

第 2 ドラフトからの変更点の概要：このアセスメントは、施策 II.11.1 に基づく独立した外部リスクアセスメントだけでなく、内部システミック リスクアセスメントのガイダンスとしても機能するため、大幅に拡張された。特に、モデル評価チームに要求される実務経験の規準を追加し、すでに高い業績を上げている研 究チームに悪影響を与えることなく、あらゆる規模の署名者に適合することを目指した。モデルへのアクセスに関する詳細を明確にし、適切なセキュリティー対策を考慮しつつ、厳密に必要な場合にのみ アクセスを増やすべきであることをより明確にした。また、「十分な時間」とは何を意味するのか、具体的な指針を示すとともに、比例原則を再度強調した。

施策 II.4.12. 安全マージン

署名者は、(1) GPAISR に起因するシステミック・リスクに適切であり、(2) (a)潜在的なエリシテーション不足、(b)システミック・リスクの評価および緩和結果における一般的な不確実性、(c)類似の評価における過去の(中)精度を代表するものであり、(3)モデルが市場で入手可能となった後の潜在的なモデル改善を考慮し、(4)入手可能な場合には、関連する最先端の手法および実務に従ったものである、十分な幅の安全余裕を特定し、実施するものとする。

本施策を実施するために、署名者は以下のことを行う：

(1) モデル評価方法が適切に較正されるよう、最善の努力を払う（すなわち、モデル評価方法が適切に較正されるよう、最善の努力を払う）。

不確かさの指標が低い場合）、可能であれば定量的な安全裕度を使用する（例えば、安全裕度を相対的、パーセンテージ、比例的な用語で表現する）。

(2) これが不可能な場合は、定性的な安全裕度を用いるか、安全裕度に対する他の適切なアプローチを採用する。例えば、マイルストーンを保守的に設定したり、(a)誤用のケースにおける潜在的または実際の敵対者、(b)合理的に予見可能なモデルの改善を考慮し、適宜、エリキューテーションの努力に資源を投入したりする。

第 2 ドラフトからの変更点の概要：ここでは主に、安全マージンの設定はシステミックリスクのアセスメントと緩和にとって極めて重要であるが、これはまだ発展途上の慣行であることを明確にした。そのため、(1)測定値の不確実性を考慮する必要があること、(2)適切な場合には、定性的または他のアプローチで行うことができることを強調した。

施策 II.4.13. 予測のための評価手法

署名者は、特に II.1.3 項を支援するため、最高システミックリスク層の中心となるモデル評価において、予測に特化 した手法を採用するよう、最善の努力を払うものとする：

(1) モデルサイズ、訓練計算量、推論計算量のスケーリング則実験、

(2) モデル世代間の比較、または

(3) 将来のモデルにおいて、危険な能力、傾向、アフォーダンス、効果がいつ出現するかの推定をサポートするその他の技術；

ただし、そのような技術は、利用可能な場合、関連する最新技術に従ったものである。

第 2 ドラフトからの変更点の概要：この措置が措置 II.1.3 と重複しないようにし、また、この分野における進展を考慮して、「スケーリング法」以外の予測技術の選択肢をさらに広げるようにした。

施策 II.4.14. 市販後モニタリング

署名者は、システミックリスクの評価と緩和に必要な情報を収集するため、市販後モニタリングを実施する。こ れには、評価するシステミックリスクに適切なモデルの能力、傾向、アフォーダンス、効果に関する情報を含む。署名者は以下を実施する：(1) 特に、GPAISR が、システミック・リスク許容規準で定義される許容できないシステミック・リスクを 引き起こさないことを確認する観点から、

市販後モニタリングを実施する（II.1.2 項）。

本施策を履行するため、署名者は、 、関連する市販後情報を収集するため、自 国の GPAISR の具体的な統合、公表、配布戦略、利用方法に最も適した組織的、契約的、技術的、またはそ の他の方法を確立し、その一方で、情報が EU 法に則って取り扱われるようにする。

署名者は、市販後のモニタリングについて、以下のような様々な適切な方法を検討するものとする：

(1) エンドユーザーからのフィードバックを収集する；

(2) （匿名での）報告チャンネルがある

(3) （重大な）インシデント報告書

(4) バグバウンティ

(5) コミュニティ主導のモデル評価と公開リーダーボード；

(6) 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使用を識別したり、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使用パターンを探したりする

(7) 学術界、市民社会、規制当局、および／または独立した研究者と協力し、GPAISR の新たなリスク、能力、連邦における影響に関する科学的研究を支援する

(8) 施策 II.4.8 に従った参加型手法などを用いて、影響を受けるステークホルダーと頻繁にダイアログを実施する

(9) GPAISR の個々のインスタンス、または GPAISR から構築された AI システム、特に AI エージェントのような自律型 AI システムのプライバシーを維持した監視と分析を支援する新しい技術的手法の開発に投資する。

(10) GPAISR のリリースおよび配布戦略を考慮し、技術的、法的、および商業的に実行可能な場合、プライバシーを保護するロギングおよびメタデータ分析方法を実施する。

(2)クローズド・ソースの GPAISR については、システミック・リスクのアセスメントと緩和 に関連するが、隠れた思考連鎖のような第三者には透明でないモデルの側面。

署名者自身が、独自の GPAISR を組み込んだ AI システムを提供/展開する場合、署名者は、当該モ デルに起因するシステミックリスクを効果的にアセスメントし緩和するために必要であれば、システムの一部として、当該モ デルを監視するものとする。

さらに、特に上記(1)から(10)の方法で、GPAISR に起因するシステミックリスクを効果的にアセスメントし緩和するために必要な市販後情報が得られない場合、署名者は、ライセンシー、川下プロバイダ、エンドユーザと協力し、これらの関係者からそのような情報を受け取り、これらの関係者から共有された関連情報を考慮するために最善の努力を払うものとする。ただし、かかる情報は EU 法に従って取り扱われるものとする。エンドユーザーが消費者である場合、署名者は、市販後モニタリングに関連する情報（ ）を共有することをオプトインする選択肢を提供するものとする。ただし、かかる情報は EU 法に従って扱われるものとする。

将来、前文に記載される可能性がある：署名者は、例えば、データを含む、署名者によるシステミックリスクのアセスメントおよび/または緩和に関連する機密情報を考慮した上で、関連情報を共有する契約をライセンシーと締結することができる。署名者は、*AI* 事務局に対し、ライセンシー、川下プロバイダ、エンドユーザとの協力や情報共有を促進するよう要請することができる。

第 2 ドラフトからの変更点の概要：ここでの主な変更履歴は、ロギングやライセンシングのような特定の技術を使用する一般的な要件はないことをさらに明確にしたことである。その代わりに、プライバシー保護などの連邦法を常に考慮に入れながら、署名者の能力、ビジネスモデル、システミック・リスクに応じて、適切な手法を選択することである。また、KPI は柔軟性に欠け、仕様の賭博に脆弱性があるため、削除した。

コミットメント II.5. システミック・リスクの受け入れ決定

法律文書：AI 第 55 条第 1 項および第 56 条第 2 項（d

署名者は、自国の GPAISR に起因するシステミック・リスクの許容度を決定することを約束する。

GPAISR のシステミックリスクとその緩和リスクとの間の比例関係を確保するため、システミックリ スク分析結果（コミットメント II.4 に従い）を、事前に定義したシステミックリスク許容規準（施策 II.1.2 に従い） と比較する。署名者は、このコミットメントに係る施策に規定されるとおり、GPAISR の開発、市場での入手 可能性、および／または利用を進めるか否かの判断材料として、この比較を用いることを約束する。

コミットメント *II.5.*を達成するために、以下を行う：

施策 II.5.1. システミック・リスクが許容可能と判断される場合の手続き

システミックリスクが許容範囲内とみなされる場合、署名者は、GPAISR の開発、最初の、または市場での入手 可能な状態、および/または利用を進める前に、残存するシステミックリスクの緩和が必要かどうかを判断する。

施策 II.5.2. システミック・リスクが容認できないと判断される場合には、手続きを行わない。

システミック・リスクが容認できないと判断された場合、署名者は、その枠組みに従い、そのリスクを容認可能なレベル まで下げるための適切な施策を講じるものとする。

モデルがまだ市場で入手可能でない場合、署名者は、例えば次のことを行う。(1) システミックリスクが許容可能と判断されるまで、追加緩和を実施する。モデルが既に市場で入手可能な場合、署名者は以下を行う：(3)モデルの市販を制限、撤回、回収する、(4)システミック・リスクに対処する。最低限、署名者は、枠組みで説明されている関連ステップに従う。

II.1.2）を参照されたい。

システミックリスクを許容可能な水準に抑えるための適切な施策が実施された場合、署名者は、GPAISR の開発、市場での最初の入手、または継続的な利用を進める前に、システミックリスク分析（コミットメント II.4 に従い）、システミックリスク受入判断（コミットメント II.5 に従い）を再度実施するものとする。

施策 II.5.3. モデル開発を段階的に進め、進捗に応じて市場で入手可能にする。

署名者は、GPAISR の開発、最初の、または継続的な市場での入手可能性、および/または使用を進める際、 必要に応じて段階的に進め、システミックリスクを許容できないレベル以下に抑え、そのレベル以下に適切に緩和するため に、以下のような最善の努力を払うものとする：(2) 市販後のモニタリングと継続的なシステミックリスクアセスメントに基づき、段階的にアクセスを拡

大す る。

第 2 ドラフトからの変更点の概要：システミックリスク評価との混同を避けるため、このコミットメントの名称を「システミックリスク受容判断」に変更した。さらに、リスク水準とシステミック・リスク受容規準を比較し、その比較に基づいて開発・展開の決定を行うというプロセスをより合理化するため、この施策と旧コミットメント 13 の「開発・展開の決定」を 1 つにまとめた。

GPAISR プロバイダの技術的リスク緩和

コミットメント II.6. 安全緩和

法律文書：第 55 条第 1 項、第 56 条第 2 項（d）、AI 法第 114 条前文

署名者は、この約束のための施策に明記されている通り、以下を約束する：(1) 技術的な

本基準に従い、当該モデルのシステミック・リスクを許容可能なレベルまで低減し、適切な 場合にはシステミック・リスクをさらに低減するために、GPAISR の開発、市場での入手 可能性、及び／又は使用から生じるシステミック・リスクに見合った、モデルのライフサイク ル全体にわたる安全緩和を行うこと。

コミットメント *II.6.*を達成するために、以下を行う：

施策 II.6.1. 実施すべき安全緩和策

II.6.2 項を損なうことなく、システミックリスクを緩和し、（本規程に従い）許容可能なレベルま で低減するために必要な場合、署名者は GPAISR に対し、以下のような技術的な安全緩和を実施する ものとする：(1) 学習データのフィルタリングとクリーニング、(2) 当該モデルの入力と出力の監視とフィルタリング、(3) 特定の要請を拒否するようにモデルを微調整するなど、安全性の観点から当該モデルの動作を変更する、(4) モデルへのアクセスを吟味されたユーザに制限するなど、当該モデルの市場での利用を制限する；(7) AI エージェントの安全なエコシステムを促進するのに役立つインフラ（レピュテーションシステム、特殊なコミュニケーションプロトコル、インシデント監視ツールなど）を実装する。

施策 II.6.2. 比例した最新の安全緩和策

署名者は、以下のような最先端の技術的安全緩和を実施するものとする：(1) (II.1.2 項に従い枠組みで規定される)システミックリスクの許容基準を考慮に入れ、問題となっているシステミックリスクに見合った緩和を行う。

この規程の目的上、最先端の技術的安全緩和が必ずしもすべてのシステミック・リスクを最大限に緩和する必要はない。

施策 II.6.3. 重大インシデント対応準備態勢

署名者は、深刻なインシデントに対処するため、（本公約および/または公約 II.7 に従い）安全および/またはセキュリティの技術的緩和を強化または追加し、是正措置を実施する。

第 2 ドラフトからの変更点の概要：現実のインシデントの複雑さに関するフィードバックに応え、II.6.3 項の是正施策を事

前に定義する要件を削除した。また、第 II.6.1 項において、署名者が考慮すべき安全緩和策を追加した（オプション）。

コミットメント II.7. セキュリティ緩和

法律文書：第 55 条 1 項、第 56 条 2 項(d)、リサイタル 114 およびリサイタル 115 AI 法

署名者は、未公開文書への不正アクセスから生じうるシステミック・リスクを緩和することを約束する。

GPAISR のモデルウェイトおよび／または未公開の関連アセット。関連資産には、アルゴリズムの洞察、訓練データ、訓練規範など、モデルの訓練に重要な情報が含まれる。

その結果、署名者は、人間や AI システムからの内部脅威を含め、十分な資金と動機づけを持つ国家レベル以外の敵対者による不正アクセスを阻止するために、最先端のセキュリティ緩和策を導入し、少なくともランド研究所 SL3 のセキュリティ目標、またはそれと同等の目標を達成し、さらに高いセキュリティ目標（例えばランド研究所 SL4 や SL5）を達成することを約束する。

コミットメント *II.7.*を達成するために、以下を行う：

施策 II.7.1. 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、少なくともランド研究所 SL3 のセキュリティ目標を満たすよう、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実施する。そのようなベストプラクティスの例としては、以下が挙げられる：

(1) 強力な ID およびアクセス管理の実践；

(2) ソーシャル・エンジニアリングに対する強力な保護；

(3) ワイヤレスネットワークの保護；

(4) 信頼できないリムーバブルメディアに対するポリシー；

(5) 物理的侵入から施設を保護する。

(6) 定期的なソフトウェア更新とパッチ管理。

本施策に基づき実施される一般的なサイバーセキュリティのベストプラクティスは、以下の事項に合致するものとする：(1) ISO/IEC 27001、NIST 800-53、SOC 2、および可能な場合はサイバーセキュリティに関する整合標準などの最新のサイバーセキュリティ技術標準、(2) NIS2 指令やサイバーレジリエンス法などのその他の関連する EU 法（適用可能な範囲）。

施策 II.7.2. セキュリティ保証

署名者は、少なくともランド研究所 SL3 のセキュリティ目標を達成するため、潜在的・現実的な敵対者に対するセ キュリティ態勢を保証し、テストするためのセキュリティ対策を実施する。このようなセキュリティ対策の例としては、以下が挙げられる：

(1) システム上のリスクを緩和するため、必要に応じて、独立した外部当事者による定期的なセキュリティレビューを行う；

(2) システミック・リスクを緩和するため、必要に応じて頻繁にレッドチームを編成する；

(3) 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；

(4) システムリスクを緩和するために必要なセキュリティテストへの一般参加を奨励するために、競争的バグ報奨金プログラムを実施する；

(5) エンドポイント検知・応答（EDR）および／または侵入検知システム（IDS）ツールを、会社のすべてのデバイスおよびネットワーク・コンポーネントにインストールする。

(6) EDR アラートを監視し、セキュリティ侵害のインシデントハンドリング、対応、復旧をタイムリーに行うセキュリティチームを活用する。

本施策に従って実施されるセキュリティ対策は、NIST 800-53（CA-2(1), CA-8(2), RA-5(11), IR-4(14)）、NIST SP 800-115 5.2 などの最新のセキュリティ保証技術標準に合致するものとする。

施策 II.7.3. 未発表モデルウェイトと関連資産の保護

署名者は、少なくともランド研究所 SL3 のセキュリティ目標を満たすよう、未公開モデル重量と関連資産を保護す るためのセキュリティ対策を実施するものとする。そのようなセキュリティ対策の例示は以下の通りである：

(1) 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安全な内部レジストリ

(2) モデルウェイトを保存するすべてのデバイスのアクセス制御と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

(3) データ、規範、サービスのみをホストする専用デバイスに保存され、モデルウェイトと関連資産を保存するデバイスに適用されるのと同等の機密性とセキュリティレベルで扱われる

(4) 少なくとも 256 ビットのセキュリティで暗号化され、暗号鍵が信頼されたプラットフォーム・モジュール （TPM）に安全に保管されるなど、最新のベスト・プラクティスのセキュリティ標準に従って、あるいは規範の発行 後にベスト・プラクティスとして確立されたより高いセキュリティ標準に従って、モデル分銅が保管と輸送 において常に暗号化されることを保証すること；

(5) モデルのウェイトが、非永続的なメモリで合法的に使用される場合にのみ復号化されることを保証する；

(6) 使用中のモデル分銅への無許可アクセスを防止するために，ハードウェアベースの信頼可能な実行 環境を使用し，最新の業界標準に従って実行可能であれば，機密コンピューティングを実施する。

(7) データセンターや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する。また、そのような場所に無許可の人員や機器がないか、定期的に検査する。

本施策に従って実施されるセキュリティ対策は、NIST 800-53 などの最新の技術標準に準拠するものとする。

施策 II.7.4. 未発表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、少なくともランド研究所 SL3 のセキュリティ目標を満たすよう、使用中のインターフェイスを強化し、 未公開モデル重みへのアクセスを制限する対策を実施するものとする。そのようなインターフェイスおよびアクセス管理施策の例示は以下の通りである：

(1) 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担当者のみを明示的に認証し、多要素認証メカニズムを通じて実施し、少なくとも 6 カ月ごとに定期的に監査

(2) 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又はデータ漏えいをセ キュリティチームが徹底的に検証する。

(3) モデルのウェイトにアクセスするインターフェースをハード化し、出力レート制限などの方法を用いて、データやウェイトの流出リスクを低減する。

本施策に従って実施されるインタフェースおよびアクセス管理施策は、NIST SP 800-171、INCITS 359-2004、NIST 800-53 などの最新の技術標準に準拠するものとする。

施策 II.7.5. 内部脅威

署名者は、GPAISR による自己潜入や妨害行為などの脅威を含め、少なくともランド研究所が定める SL3 のセキュリティ目標を満たすよう、内部脅威を選別し、保護するための対策を実施する。そのような対策の例示は以下の通りである：

(1) 未発売のモデルの重さ、関連する資産、またはそのような未発売のモデルの重さの保管や使用を制御するシステムにアクセスできる、またはアクセスする可能性のある従業員や請負業者の身元調査を行う

(2) 内部脅威の認識と報告に関する研修の提供、あるいは（3）GPAISR のサンドボックス化である。

本施策に従って内部脅威を選別し、保護するために実施される保護策は、NIST800-53（PM-12、PS-3）などの最新の技術標準に合致したものでなければならない。

施策 II.7.6. 適用範囲

署名者は、第 II.7.1 項～第 II.7.5 項に規定されるセキュリティ対策が、(1)訓練前から、安全な削除、またはモデルウェイトや関連資産 を公表する決定までのモデルライフサイクル全体に適用されること、(2)(1)項を損なうことなく、GPAISR に起因するシステミックリ スクに従い、セキュリティ対策の実施に優先順位をつけること、を確実にするものとする。

施策 II.7.7. 限定的開示

署名者は、一般に入手可能な枠組みまたはモデル報告書の写しが、II.7.1-II.7.5 項に従って検討、実施されたセキュリティ緩和策を開示することを保証するものとする、技術的な安全緩和の有効性を損なうこと（ コミットメント II.6 に準拠）、または公表の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を漏洩することなどによって生じ得る、システミックリスクの実質的な増大を防止するために必要な場合には、再編集を行うこと。

施策 II.7.8. より高いセキュリティ目標

ランド研究所 OC4 レベルの敵対者によるセキュリティ上の脅威が合理的に予測される場合、署名者は、少なくともランド研究所 SL4 のセキュリティ目標またはそれに相当するものに沿って、より厳格なセキュリティ緩和策の研究を進め、実施するものとする。

第 2 ドラフトからの変更点の概要：セキュリティ対策の規定性についてのフィードバックを受け、特定のセキュリティ対策ではなく、ランド研究所 SL3 のセキュリティ目標に焦点を当てた成果ベースのアプローチを採用した。新しい「約束 II.7」の各「対策」は、特定の手段や技術を規定することなく、署名者がどのようにセキュリティ目標を達成するかについて、より柔軟性を持たせた。

また、不正な内部展開の脅威モデルを考慮するため、施策 II.7.5 において内部脅威としての GPAISR への言及を追加した。

最後に、II.7.6 項では、セキュリティ施策の比例的な適用に関する意見に応え、より大きなリスクをもたらす可能性のある GPAISR、例えば、早いチェックポイントよりも遅いチェックポイントを優先する柔軟性を認めた。

GPAISR プロバイダのガバナンスリスク緩和

コミットメント II.8. 安全・安心モデルレポート

法律文書：第 55 条第 1 項、第 56 条第 2 項（d）および第 56 条第 5 項 AI

署名者は、実務規範の実施状況、特に以下の事項を AI 事務局に報告することを約束する。

を作成することにより、GPAISR の開発、市場での入手可能性、及び／又は使用に自らの枠組みを適用する。この枠組みは、このコミットメントのための施策に規定されているとおり、GPAISR を市場で入手可能にするごとに、安全・セキュリ ティモデル報告書（以下、「モデル報告書」という。

コミットメント *II.8.*を達成するために、以下を行う：

施策 II.8.1. 詳細レベル

署名者は、モデル報告書に以下の詳細レベルを記載する：

(1) 当該モデルがシステミック・リスクのレベルに達しているか、または当該モデルがモデル・ライフ サイクル全体を通じてシステミック・リスクのレベルに達することを署名者が合理的に予見していることに比例し、かつ、署名者がシステミック・リスクの評価と緩和策を規程に従ってどのように実施しているかをアセスメント事務局が理解できるようにすること。

モデル報告書には、(1)および(2)を満たす十分な理由が記載されていなければならない。

将来の前文への資料となる可能性がある：*(1)*の下で、より低い詳細レベルを正当化しうる理由付けの例として、モデル報告書は、そのモデルが *II.4.2* 項の意味において安全に導き出されたモデルであることを主張することもできるし、そうでなければ、*EU* 市場で入手可能な既存のモデルと類似していることを主張することもできる

施策 II.8.2. 当該モデルを市場で入手可能とすることを決定した理由

署名者は、モデル報告書に、当該モデルを初めて市場で入手可能とすることを含め、その決定を正当化するた めに使用した理由と情報を記載する：

(1) GPAISR のシステミックリスクが II.5.1 に従い許容可能であることを、II.4.12 項に従った十分な安全マージン及びそれを評価するための説明を含め、II.5.1 項に従った十分な根拠を提供する、モデル報告書に提供された情報及びデータに基づく明確な推論の連鎖

(2) (1)の署名者の推論が成り立たなくなる条件、及び、コミットメント II.11 に従ったモデルの評価などを通じて、独立した外部評価者が、 のモデルを市場で入手可能にする、または使用する決定を通知したかどうか、またどのように通知したか。

施策 II.8.3. システミックリスクのアセスメントと緩和の文書化

モデル報告書において、署名者は以下のことを行う：

(1) また、必要に応じて附属書 1.2、1.3 及び 1.4 を参照しつつ、GPAISR の影響度の高い能力に特有な他のシステミック・リスクを選択するための努力の説明を含む

(2) (II.4.1 項に従って)実施されたシステミック・リスクの見積りプロセスを文書化し説明し、当該プロセスの一環として行われたすべての決定、特に定性的及び／又は定量的な見積りの選択、並びに当該見積りがシステミック・リスクの階層とどのように関連しているかを正当化する根拠を提供する

(3) 施策 II.4.2 に従って「安全な派生モデル」の概念を使用する場合は、「安全なオリジネー ターモデル」の規準（施策 II.4.2、ポイント(1)による）及び「安全な派生モデル」の規準（施策 II.4.2、 ポイント(2)及びポイント(3)による）がどのように満たされるかについて、その正当性を示す

(4) 特に、(a)（II.4.12.に従った）安全裕度、(b)採用したモデル評価の厳密性の低いレベル、(c)そのような低いレベルを正当化する理由、及び(d)実施したモデル評価の結果；

(5) GPAISR のシステミックリスクの評価に使用した各モデル評価について、そのモデル抽出作業を文書化し、説明する（施策Ⅱ.4.6 に準拠）

(6) 実施した各モデル評価について、施策Ⅱ.4.7 に従って AI システム情報を考慮した方法を文書化し、説明し、当該モデル評価が当該システミック・リスクの効果的なアセスメントにとって適切であった理由を説明する

(7) (b)（II.11.1 項に従って）市場でモデルを入手可能にする前に実施されるアセスメントに関 与する独立した外部アセスメントについても、(a)で要求される情報と同様の情報を提供する。(b)に関しては、署名者がモデル報告書を作成するのと同時に、そのような独立した外部評価者に対し て、必要な情報を AI 事務局に直接提供するよう要請する；

(8) 当該モデルの評価されたシステミック・リスクを、（コミットメントⅡ.6 及びⅡ.7 に従って実施された）安全・セキュリティ緩和の有無にかかわらず、適宜比較し、説明する

(9) 当該モデルについて実施されたすべてのシステミック・リスク緩和策を文書化し、説明し、その限界を

さらに、非 SME 署名者のみが、これを行うものとする：

(10) システミック・リスクを評価するために使用された各モデル評価について、当該モデル評価が最新のものであり、かつ、施策 II.4.4 の(1)から(3)を満たすと結論付けた根拠を文書化し、説明する。

(11) 使用した各モデル評価について、内部妥当性、外部妥当性、再現性を文書化し、説明する（施策Ⅱ.4.5 に従う）；

(12) 使用された各モデル評価について、そのモデル評価（施策 II.4.8 に準拠）が扱う想定される使用状況および様式を文書化し、説明する

(13) また、（施策 II.4.9 に従って）モデル評価のためのどのツール、ベストプラクティス、および／またはデータを誰と共有したか（施策 II.4.10 に従って）をモデル報告書に記載するために最善の努力を払う。

施策 II.8.4. 外部レポート

署名者は、模範報告書に記載する：

(1) (a)当該モデルが市場で入手可能となる前に当該モデルをレビューした独立した外部評価者（施策Ⅱ.11.1 に従う）、及び(b)独立した外部当事者により実施されたセキュリティレビュー（施策Ⅱ.7.2 の(1)に従う）から、既存の守秘義務（商業上の守秘義務を含む）を尊重し、かつ、当該外部評価者又は当事者が所見の公表について制御を維持できる範囲において、入手可能な報告書（例えば、妥当性確認によるハイパーリンク）を提供する；

(2) 独立した外部評価者がモデル評価またはシステミック・リスク評価に関与していない場合には、第 II.11.1 項に従った当該関与が必要な規準が満たされていない理由、または第 II.11.1 項に従って独立した外部評価者を特定し選定するための最善の努力を行ったにもかかわらず、独立した外部評価者が見つからなかった理由の説明、および、少なくとも 1 名の独立した外部評価者がモデル評価またはシステミック・リスク評価に関与している場合には、当該評価者が関与した後に AI オフィスにより承認された場合を除き、第 II.11.1 項の資格規準に基づく、評価者の選定の正当性の理由。

施策 II.8.5. アルゴリズムの改善

署名者は、モデル報告書に、当該 GPAISR に特有のアルゴリズムまたはその他の改善点に関し、既に市販されている他の GPAISR と比較し、AI 事務局がシステミック・リスクの状況の重大な変化を理解する上で有 益であり、それにより AI 事務局が本規程が求めるシステミック・リスクの評価と緩和の実際的な実施をよりよく理解できるような、すべてのハイレベルな情報が含まれていることを確認する。

施策Ⅱ.8.6 意図的なモデルの特定

署名者は、モデル報告書に、例えば以下のような、署名者がどのようにモデルを運用するつもりであるかの仕様が含まれていることを確認する：

(1) モデルが従うべき原則を規定する

(2) モデルが異なる種類の指示にどのように優先順位をつけることを意図しているかを記載する、または、

(3) 指示を拒否することを意図したトピックを列挙する。

施策 II.8.7. モデルレポートの更新

システミックリスクの状況に重大な変化が生じ、モデルが許容可能なシステミックリスクをもたらすと結論づけた 理由が著しく損なわれたと信じるに足る理由がある場合、署名者はモデル報告書を更新しなければならない（Measure II.8.2, point (1)）：

(1) モデルの能力が著しく変化した場合、それは事後訓練、さらなるエリシテーション、またはアフォーダンスの変化によるものである；

(2) データ・ドリフト、関連するモデル評価手法の大幅な改善、または以前のシステミック・リスク・アセスメントがもはや正確でないことを示唆するその他の要因である；

(3) 緩和の改善

(4) 重大インシデントである；

(5) 市販後のモニタリングから得られた情報（施策Ⅱ.4.14 に従う）、例えば、当初のアセスメントで予測された使用方法と比較して、モデルの使用方法が大幅に変更された場合など。

(6) モデルの内部使用から得た情報である。

署名者は以下を行うものとする：(1) システミック・リスクの状況を継続的に検証する。(2) モデル報告書を更新するため、システミック・リスクの状況に重大な変化が生じたと信じる理由を特定し、それに関連する情報を収集する。

署名者は、更新されたモデル報告書に少なくとも以下の情報が盛り込まれていることを確認する：

(1) 適切な場合には、前回のモデル報告書の内容および過去に実施されたシステミック・リスク・アセスメントの結果

(2) 前回のモデルレポートでは考慮されなかった、モデルに依存しない情報で

(3) 前回のモデルレポートでは考慮されなかったモデル評価である；

(4) ヒヤリ・ハットなどの関連情報を含む重大インシデント報告書（コミットメント II.12 に基づく）からの情報を、入手可能な限り提供する；

(5) 施策 II.4.14 に基づく）市販後モニタリングからの情報、およびモデルの内部使用からの情報。

更新されたモデル・レポートには、少なくとも以下の内容が含まれていなければならない：

(1) 前回のモデル報告書に記載されていないシステミック・リスク・アセスメントの関連する結果及び文書、並びにモデル報告書の更新を促した情報を含め、関連する場合には更新された前回のモデル報告書の内容

(2) II.11.2 項に従い、独立した外部アセスメント機関が提出した、署名者の GPAISR のシステミックリスク評価、テスト手法、緩和案に関連する新たな報告書

(3) 前回のモデル報告書以降、当該モデルに関して枠組みが遵守されているかどうかのアセスメント

(4) モデル報告書がどのように更新されたかを、バージョン番号と変更日とともに記述した変更履歴。

(5) 該当する場合には、本施策のプロセスに従った結果、署名者が、前回のモデル報告書と比較してシステミックリスクの状況に重大な変化がなく、その時点で当該モデルについてさらなるシステミックリスクの評価と緩和が必要ないと結論づけた理由を説明すること。

第 2 ドラフトからの変更点の概要：このコミットメントの主な変更点の一つは、モデル報告書をいつ、どのように更新すべきかを詳述し、このプロセスが例えば市販後モニタリングとどのように関連するかを明確にしたことである。第 2 ドラフトのモデルレポートコミットメントには、少なくとも 6 ヶ月ごとにモデルレポートを更新しなければならないと関係者が読み取った部分があった。現在のドラフトでは、プロバイダが「システミック・リスクの状況に重大な変化があり、モデルが許容可能なシステミック・リスクを有すると結論付けた理由を著しく損なうと信じるに足る理由がある」場合に、モデル報告書を更新する必要がある。

もうひとつの変更点は、モデル報告書で取り上げるべき情報について、これまでは規程のあちこちに散らばっていた内容を、このコミットメントに統合し、解析しやすくしたことである。

また、フィードバックで指摘されたように、特定の緩和策（例えば、セキュリティ緩和策、事前学習に組み込まれたアライメント技法）を取り除くことが必ずしも適切でない場合があるため、施策 II.8.3 における緩和前後のリスクアセスメントを「適切」のみとし

た。

施策 II.8.2 では、展開の判断材料となりうる要素が複雑であるとの意見を受け、システミックリスク緩和の十分性に焦点を当てた。

施策 II.8.4 では、外部報告書をモデル報告書自体に逐語的に含める必要はなく（例えば、参照は許容される）、そのような報告書を含める場合は、例えば公表の制限に関して、既存の合意を尊重すべきであることを明確にした。

コミットメント II.9. アセスメント

法律文書：第 55 条第 1 項、第 56 条第 2 項（d）および第 56 条第 5 項 AI

署名者は、コミットメント II.1 のもとで採択と実施を約束した枠組みの妥当性を評価し、このコミットメントに 関する施策で規定された結果に基づき、その枠組みを更新することを約束する。

コミットメント *II.9.*を達成するために、以下を行う：

施策 II.9.1. アセスメントプロセス

署名者は、アセスメントを実施する際、以下を考慮するものとする：(1)ベストプラクティス、(2)関連研究、最先端の科学、(3)モデルのインシデントや故障、重大インシデント報告、(4)既に利用可能な場合は、関連する独立した外部の専門知識。

署名者は、可能な限り速やかに、遅くとも完了後 5 営業日以内に、監督機能を担う署名者の経営組織または他の適切な独立機 関（審議会や理事会など）に対し、結果のアセスメントを提供する。

施策 II.9.2. モデル別妥当性評価

安全な派生モデル（施策 II.4.2 による）は、以下の施策の適用除外となる。

自国の GPAISR に関する枠組みの適切性を評価するため、署名者は、以下のモデル固有の情報を適切性評価に記 載し、汎用 AI モデルが AI 法第 51 条(1)(a)の分類条件を満たしている、または満たす見込みであることを AI 事務局に通知してから 4 週間以内に、II.14.3 項に従い、更新された適切性評価を AI 事務局に提出するものとする：

(1) **モデルの説明**：他の AI モデルの使用も含め、モデル開発に使用する技術や資産の一般的な説明

(2) **システミックリスクの特定**：コミットメント II.3 に従い、枠組みに従い、本モデルについて実施されたシステミック・リスク識別の結果

(3) **予測：**利用可能なすべての予測（施策Ⅱ.1.3 および施策Ⅱ.4.13 の技法に基づく）であって、特に当該モデルが意図する、または期待される能力、システミック・リスク指標、システミック・リスクの程度を示す

(4) **システミックリスク分析結果**：コミットメント II.4 に従い、入手可能な全てのシステミック・リスク分析結果；

(5) **システミックリスクのアセスメントと緩和計画**：**システミック・リスク**分析およびシステミック・リスク許容決定（コミットメント II.4-II.5）、ならびに技術的なシステミック・リスク緩和（コミットメント II.6-II.7）の概要：開発段階におけるシステミック・リスク：モデルの開発段階において、どのようなシステミック・リ スクが発生する可能性があるか（もしあるのであれば）、また、開発段階において重大なシステミック・リ スクが発生する可能性がどの程度あるかについてのリスクアセスメント。重大なシステミックリスクは、例えば、以下のような場合に開発フェーズ中に発生する可能性がある：

(a) 署名者のセキュリティ緩和が、モデルの予測システミック・リスク・プロフィールに対して不適 切である

(b) モデルは、安全緩和を実質的に損なうことが可能であり、かつ/または、モデルの流出を実行するのに十分な自律的能力を有している。

(c) 例えば、研究者の生産性が 5 倍加速されるとか、一流 AI 企業の能力研究チームが通常行っている作業を AI モデルが同程度の総費用で行っている場合などである。

特定の状況においては、署名者は、初回の妥当性評価後、II.2.2 項(1)号で定義されるマイルストーンに到達するたびに、II.14.3 項に従い、妥当性 評価を更新し、AI 事務局に提出するものとする：

(1) 重大なシステミックリスクは、上記(6)の通り、開発段階で発生する；

(2) 上記(1)～(6)のいずれかに重要な変更があった；

(3) 過去 12 週間以内に当該モデルに固有のアセスメントが完了していない。

例えば、AI の研究開発能力がモデル開発に関連するシステミックリスクの場合、AI 労働の自動化レベルや、 実施された安全・セキュリティ保護施策に重点を置き、モデルの漏洩や流出に起因するシステミックリスクの場合、 追加のセキュリティ緩和策を実施するための署名者の努力に重点を置く。

施策 II.9.3. 枠組みの妥当性評価

署名者は、（重大インシデント報告などを通じて）枠組みの適切性が著しく損なわれる可能性のある 重大な変更が認識された場合、または、GPAISR の最初の市場入手可能日から 12 カ月ごとに、いずれか早い方 法で枠組みの適切性アセスメントを実施する。

枠組みの妥当性アセスメントは、以下の質問を対象とする：

(1) **実務規範に準拠した枠組み**枠組みは、コミットメント II.1 で説明されているすべての構成要素に、その有効性を評価するのに十分な詳細さで対応しているか。その他の点で、枠組みは実務規範と整合しているか。

(2) **枠組みの適切性：**枠組みの施策、手続、資源及び緩和は、今後 12 ヶ月以内に市場で入手可能 になる予定の署名者の GPAISR に起因するシステミックリスクの複雑さ及び規模に 対して、依然として適切であるか。枠組みは、署名者のモデルが今後 12 ヶ月間に外部及び内部でどのように利用され、また 利用されると予想されるかを考慮した上で、適切であるか。枠組みは、システミックリスクのアセスメント及び緩和手法の有効性だけでなく、将 来の展開における不確実性も適切に考慮しているか。

(3) **枠組みの順守：**今後 12 カ月間、枠組みが遵守されると確信できる強い理由があるか。

第 2 ドラフトからの変更点の概要：多くの利害関係者、特にプロバイダは、第 2 ドラフトのアドヒアランスとアセスメントに関するコミットメントが、 署名者によるシステミックリスクの評価と緩和において、相応の利益が得られないまま、事務量が増えすぎるとの懸念を 表明した。この懸念は、特にアドヒアランス評価と、アセスメントを実施する頻度に関して強 く、特定の間隔で実施するのではなく、重要な変化に対応してアセスメントを実施することを強く望んだ。

これらの懸念に対応するため、プロセスを合理化した。アセスメントを削除し、その代わりに、重要な変更に対応してモデル報告書を更新するプロセスを明確化した（施策 II.8.7 参照）。さらに、新しいモデルの上市を意図した場合に発動されるアセスメントと、より一般的なフレームワークの妥当性アセスメントを分離した。後者は、署名者が関連するシステミック・リスク緩和を適切な時期に実施できることを保証するものであり（特にセキュリ ティに関する緩和の中には、実施に 1 年を要するものもあるため）、前者は、署名者が上市を意図するモデルを十 分に安全かつ確実に取り扱うことができることを保証するものである。

第 4 次ドラフトでは、こうしたプロセスをさらに合理化するために、このコミットメントをさらに洗練させることが必要になるだろう。

コミットメント II.10. システミック・リスクの責任分担

法律文書：第 55 条第 1 項、第 56 条第 2 項（d）、AI 法第 114 条前文

GPAISR のシステミックリスクのアセスメントと緩和に関する活動に関し、署名者は以下を約束する、

このコミットメントのための施策に明記されているとおり、以下を行う：(1) GPAISR からシステミック・リスクを管理するための責任（ ）を明確に定義し、組織のあらゆる階層に割り当てること、(2) システミック・リスクを管理する責任を割り当てられた行為者に適切なリソースを割り当てること、(3) 健全なリスク文化を促進すること。

署名者は、少なくとも署名者の組織の複雑さとガバナンス構造、および GPAISR に起因するシステミックリスクに 比例した、適切なレベルの責任とリソースを割り当てることを約束する。

コミットメント *II.10.*を達成するために、以下を行う：

施策 II.10.1. 責任の明確化

署名者は、GPAISR に起因するシステミック・リスクマネジメントの責任を、組織の全レベルにわたって明確に定義する。これには以下の責任が含まれる：

(1) **リスクの監督**署名者のシステミックリスクのアセスメントと緩和活動を監督する

(2) **リスクの所有：**署名者の GPAISR から生じるシステミック・リスクをマネジメントする

(3) **支援と監視：リスクアセスメント：** 署名者のシステミックリスク評価と緩和活動を支援、監視する：アセスメント： 署名者のシステミックリスク評価と緩和活動が適切であることを、監督機能を有するマネジメント機 関、または他の適切な独立機関（審議会や理事会など）に対し、内部および必要に応じて外部で保証する。

署名者は、そのガバナンス構造と組織の複雑性に応じて、これらの責任を組織の以下のレベルに割り当てる：

(1) 監督機能を有する経営層または他の適切な独立機関（審議会や理事会など）

(2) 経営層がその経営機能を果たす

(3) 関連する運営チームである

(4) 内部保証プロバイダ（内部監査機能など）（利用可能な場合）、および（5）外部保証プロバイダ（サードパーティ監査人など）（利用可能な場合）。

この施策は、署名者が順守している場合、履行されたものとみなされる：

(1) (a)ISO/IEC23894:2023（4.5.3 項）、(b)ISO31000:2018（4.5.3 項）、(c)NIST AI リスクマネジメントフレームワーク（ガバナンス 2）、又は(d)IIA のスリーラインモデル。

(2) 以下のすべてである：

(a) **リスクの監督**署名者のシステミックリスク管理活動を監督する責任は、マネジメント団体の特定の委員会（リスク委員 会や監査委員会など）に割り当てられる。中小企業である 署名者の場合、この責任は、監督機能を有する経営組織の委員会メンバー個人に 割り当てられる。

(b) **リスクの所有権** GPAISR から生じるシステミック・リスクの管理責任は、そのマネジメント機能において、リスク を生じさせる可能性のある関連する署名者の中核的事業活動にも責任を有する、適切なメンバー に割り当てられている。

研究・商品開発など、システミック・リスクに関連する業務（研究責任者や商品責任者など）を担当する。マネジメント団体のメンバーは、システミック・リスクを生み出す活動の一部（例 えば、特定の研究領域や特定の商品）を監督するオペレーション・マネジャー に、より低いレベルの責任を割り当てている。組織の複雑さによっては、責任のカスケード構造が存在する場合もある。

(c) **サポートとモニタリング**署名者のシステミック・リスク管理活動を支援し、監視する責任は、そのマネジメント機能を担う 団体のメンバー（例えば最高リスク責任者）に割り当てられている。この構成員は、システミックリスクを生み出す可能性のある加 盟者の中核的な事業活動（例：研究、商品開発）にも責任を負ってはならない。この担当者は、必要に応じて、中央リスク機能および他の業務チームから支援を受ける。中小企業である署名者の場合、署名者のシステミックリスクの評価と緩和を支援し、監視することを任務 とする個人が、少なくとも 1 名、マネジメント組織に存在する。

(d) **保証：**署名者のシステミックリスクの評価及び緩和活動の適切性について、マネジメント団体の監 督機能または他の適切な独立機関（審議会や理事会など）に保証を提供する機能は、関連者（監査責任者 や内部監査責任者など）に割り当てられている。この個人は、内部監査機能および適宜の外部保証によってサポートされる。内部監査機能は、経営機能において経営層から組織的に独立している。署名者の内部保証活動は、業界のベストプラクティス（例えば IIA の「グローバル内部監査標準」に示され ているもの）に従う。中小企業である署名者の場合、監督機能のマネジメント機関は、署名者のシステミックリスク評価と緩和 を定期的に評価する。（例えば、署名者の枠組み妥当性評価を承認するなど）。

施策 II.10.2. 適切な資源の配分

署名者は、署名者の GPAISR に起因するシステミックリスクのレベルに見合った適切な資源が、GPAISR に起因するシステミックリスクを管理する責任を割り当てられた主体（以下を含む）に配分されるよう、そのマネジメント団体が監督することを保証する：

(1) 人材

(2) 財源

(3) 情報と知識へのアクセス、そして計算資源

この施策は、署名者が以下のいずれかを順守している場合、またはそれらが置き換えられるなどして 廃止されていない場合、履行されたものとみなされる：

(1) ISO/IEC23894:2023（4.5.4 節

(2) ISO 31000:2018（4.5.4 項）；または

(3) NIST AI リスクマネジメントフレームワーク（Manage）。

施策 II.10.3. 健全なリスク文化の推進

署名者は、健全なリスク文化を促進し、（II.10.1 項に従い）GPAISR に起因するシステミックリスクのマネジメントを担 当する行為者が、システミックリスクのレベルに応じて、不適切なリスク追求、リスク無視、リスク回避のいずれでもなく、システミックリスクに対する慎重かつバランスのとれたアプローチをとることを確保するための施策を講じるものとする。

署名者の GPAISR。

本対策における健全なリスク文化の兆候とは、例えば以下のようなものである：

(1) 健全なシステミック・リスク・カルチャーに関して、トップが基調を示す

(2) システミック・リスクに関する意思決定に対する効果的なコミュニケーションと異議申し立てを可能にする

(3) 慎重な行動に対する報酬や、システミック・リスクの内部フラグを立てるなど、過度なシステミック・リスクテイクを抑制するための適切なインセンティブを設ける

(4) 匿名調査により、職員が報告手段を認識し、システミックリスクについて懸念を表明しやすく、 署名者の枠組みを理解し、安心して発言できることが判明する。

(5) 内部報告チャネルは積極的に利用され、報告は適切に処理される。

この施策は、署名者が以下のいずれかを順守している場合、（それらが代替されていない、あるいは陳腐化していない場合）、履行されているとみなされる：

(1) ISO/IEC23894:2023 と ISO31000:2018（第 4、5.2～5.4、6.1 項）を併用する。

(2) NIST AI リスクマネジメントフレームワーク（ガバナンス 4）。

第 2 ドラフトからの変更点の概要：一部の利害関係者は、第 2 ドラフトのシステミックリスク責任配分のコミットメントが規定的すぎるとの懸念を表明した。これは、一部の利害関係者が KPI を強制的なものであると読んだことが一因であり、ドラフトの意図とは異なるものであった。また、コンプライアンスを証明する手段として、既存のどの標準を使用できるかを明示してほしいという要望もあった。私たちは、コミットメントが達成すべき全体的な目標を明示した上で、関連標準の遵守を含め、署名者が遵守を証

明できる具体的な方法（排他的ではない）を説明することで、これら 2 つの懸念に対処した。

また、適切な場合には、関連する理事会や審議会が監督機能において経営組織の役割を果たすことができることを明確にするなど、コミットメントがさまざまな規模やタイプの組織（ ）に適したものとなるよう努めた。コミットメントが、まだ異なる組織形態を適切に考慮できていない可能性がある点についてのフィードバックを歓迎する。

コミットメント II.11. 独立した外部評価者

法律文書：第 55 条 1 項および AI 法第 114 条前文

署名者は、II.11.1 項および II.11.2 項に規定される範囲および条件のもと、モデルライフサイクル全体にわたり、 GPAISR のモデル評価を含む独立した外部システミックリスクアセスメントを取得することを約束する。

コミットメント *II.11.*を達成するために、以下を行う：

施策 II.11.1. 市場導入前のアセスメント

2025 年 11 月 2 日以降、モデルを市場で入手可能とする前に、署名者は、コミットメント II.3 に従って特定されたすべてのシステミック・リスクについて、その GPAISR が独立した外部評価者による、当該特定のシステミック・リスクに関連するモデル評価を含むシステミック・リスクアセスメントを受けることを保証する：

(1) が市場で入手可能となった；

(2) は、当規範に従ったシステミック・リスク・マネジメント・プロセスを経て、関連署名 国のシステミック・リスク受入規準を満たすか、2025 年 5 月 2 日より前に市場で入手可能になっている；

(3) 未発生の重大インシデントによって特定されるような、重大な安全性またはセキュリティ上の欠陥があることが示されていない。

(4) 署名者にとって十分な透明性がある(署名者が、安全緩和、セキュリティ緩和、アーキテ クチャ、能力、アフォーダンス、モダリティ、システミックリスクプロファイルなど、モデルの安全性に関 連する特性について十分な可視性を持つことを意味する。)

本測定における同程度の安全性の規準には、他のモデルとの比較が含まれる：

(1) 署名者は、コミットメント II.3 に従ってシステミック・リスクを特定した後、そのモデルについて新たな、あるいは異なるシステミック・リスク・シナリオを合理的に予見していない

(2) 関連する能力と識別されたシステミック・リスクを測定する最新のベンチマーク（コミットメント II.3 による）、または署名者のモデルに関する他の最新のアセスメントにおけるスコアがすべて低いか同等であり、かつ署名者が II.4.11 の(1)に従い十分な資格を確保している

(3) (4)署名者のモデルが、より大きなシステミック・リスクをもたらすと考える理由が他にない。

署名者は、早期の調査、少なくとも 2 週間前からの透明性のある公募、適切な評価者の早期通知などを通じて、 (1)AI 事務局が認めた、または(2)独立した外部評価者を特定し、選択するために最善の努力を払うものとする：

(a) 評価対象であるシステミック・リスクの領域に関する重要な専門知識を有し、かつ、規準 II.4.11 の(1)に従い、関連する場合には、モデル評価の実施に技術的な熟練と経験を有している、

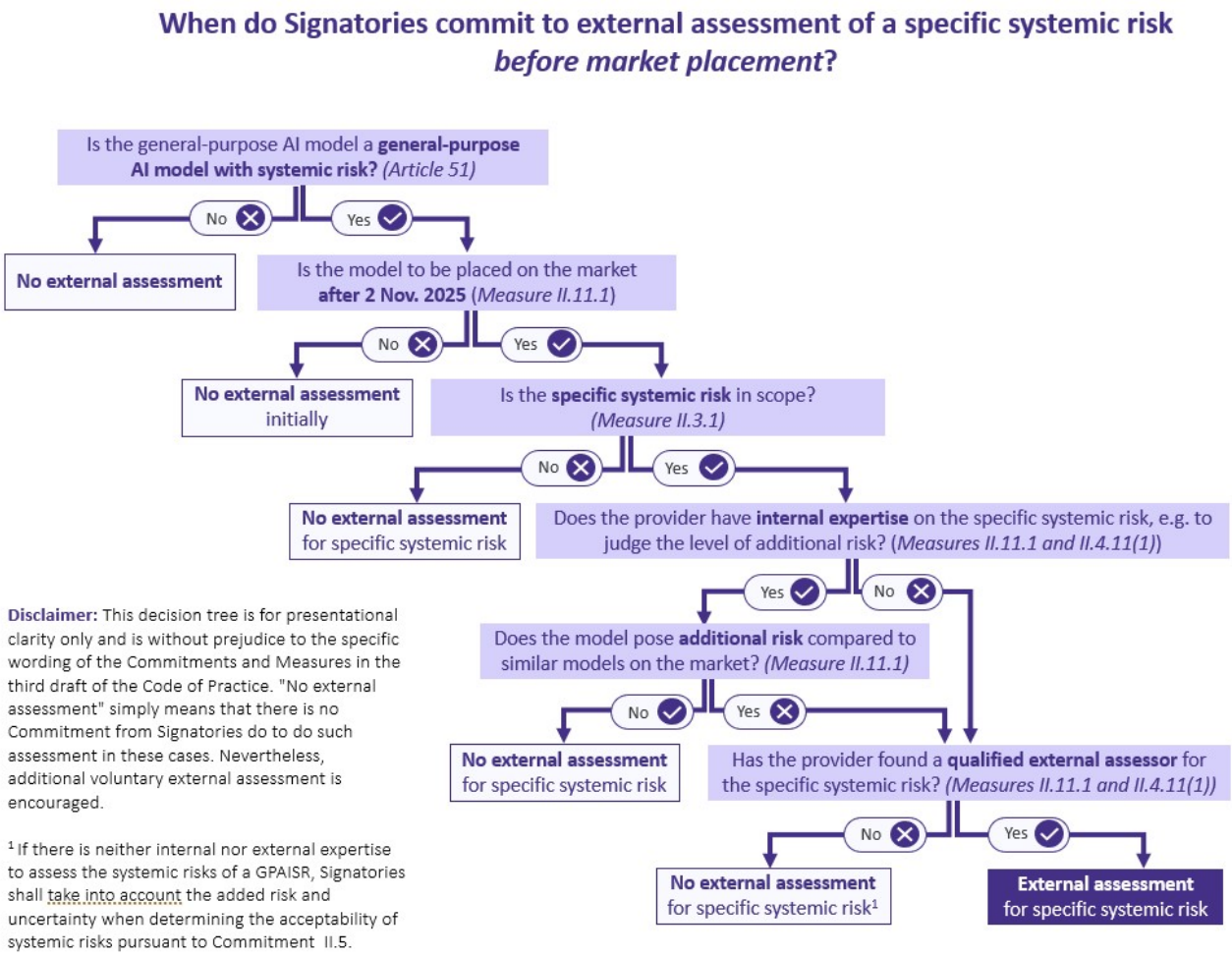
ただし、独立した外部評価者がそのような情報へのアクセスを必要とする場合、商業上の機密情報を保護することに同意していることを条件とする。

署名者は、システミック・リスクを効果的に評価するために必要な範囲において、独立した外部アセスメント担当 者に対し、十分なアクセス、情報、時間、リソースを提供する。

II.4.11.署名者は、明示的な許可なく試験実施から得られたモデルの入力と出力を保存、分析することで、外部アセスメ ントの完全性を損なってはならず、適切な場合には、独立した外部評価者に安全な推論を提供するものとする。

署名者のモデルが上記で規定された安全な派生モデルでも同様の安全モデルでもないにもかかわらず、II.11.1 項の規準に従った資格を有する独立した外部アセスメント機関を特定できなかった場合、署名者は、コミットメント II.5 項に従いシステミックリスクの許容性を判断する際、外部アセスメントの不在から生じるシステミックリスクの潜在的な追加リスクと不確実性を考慮するものとする。

将来、前文に記載される可能性がある：そのモデルが、既に市場に存在する *GPAISR* に起因する以上のシステミック・リスクをもたらさない場合、 署名者は、システミック・リスク評価プロセスに外部のアセスメント機関を関与させることで、より専門的な知 識を導入し、システミック・リスク評価について独立した保証を提供できることを認識する。



施策 II.11.2. 市場導入後のアセスメント

(1)市場で入手可能な状態のモデル、(2)参考モデルや基本モデルなど、特定の緩和を伴わないモデル。署名者は、システミックリスクに関連する非商業的かつ合法的な研究を実施する評価者に、当該研究プロ グラムのための無料の研究 API クレジットを割り当てるものとする。署名者は、明示的な許可なくテスト実行のモデルインプットとアウトプットを保存・分析することで、外部アセスメ ントの完全性を損なわないものとし、適切な場合には、独立した外部評価者に安全な推論を提供する。

本施策の独立した外部評価者を特定し選択するため、署名者は、独立した外部評価者からの申請を評価 するための、関連する業務経験や学問的経験を反映した明確かつ透明性のある規準を公表し、適切な場 合には、これらのモデルの感度が高くなっていることを考慮し、特定の緩和施策のないモデルへのアクセスに ついては、より厳しい規準とする。

署名者は、アセスメントの結果として、アセスメント実施者に対していかなる法的・技術的報復も行わないこと、また、該当する場合には、アセスメント実施者が調査結果を公表することを含め、探索的な独立外部アセスメントのための法的・技術的セーフハーバー体制に貢献するものとする：

(1) 明示的に許可されている場合を除き、テストを通じて意図的にシステムの可用性を中断させないこと

(2) 査定者は、利用者の明示的な同意なしに、機微または機密の利用者データに故意にアクセス、修正、および/または使用しない。査定者がかかるデータにアクセスする場合は、報告に必要なもののみを収集し、直ちに報告し、拡散を控え、法的に可能な限り速やかに削除しなければならない

(3) ただし、事前に合意された方針とタイムラインの下での公開は、そのような強要とみなされない。 責任ある脆弱性公開のために、署名者が公開するプロセスを遵守する。

GPAISR を完全にオープンソース化することは、この「施策」を満たす代替手段を提供する。

第 2 ドラフトからの変更点の概要：多くの利害関係者は、外部アセスメントに関連する専門知識、十分なアクセス、十分な時間に関して、第 2 ドラフトで提供された詳細の欠如を懸念していた。これらの点については、内部審査と外部審査の両方に関連性があり、また規程全体を通しての繰り返しを減らすために、第 II.4.11 項において主に取り上げられていることに留意されたい。

また、ステークホルダーからは、(a)市場導入前に外部アセスメントが必要でない場合の 1 つの状況として第 2 ドラフトで提示した類似性の概念と、(b)そのような類似性の「十分な証拠」を構成するものについても明確にしてほしいとの要望があった。我々は、このドラフトで明確化を図った。一部の利害関係者からのフィードバックに基づき、外部アセスメントが不要となる可能性のあるもう一つの状況として、「安全に導き出されたモデル」（施策 II.4.2 で定義）の概念を追加した。

現在の評価エコシステムは、外部アセスメントをサポートするほど成熟していないという懸念を、関係者から引き続き受けていた。これらの懸念を軽減するために、我々は 11 月までの猶予期間を設けた。

2025 年、施策 II.11.1 に従い、GPAISR の適切な外部評価者を見つける。加えて、企業秘密を尊重し、商業上の機密情報を保護することに同意しない外部評価者と協働しなければならないことを懸念する関係者もいた。我々は本ドラフトにおいて、プロバイダは、たとえそれが II.11.1 項に従って適格な外部評価者が特定できないことを意味するとしても、このような理由で評価者との協働を拒否できることを明確にした。

また、前回のドラフトでは、AI 事務局が潜在的な評価者として名前が挙げられていたため、ステークホルダーから懸念の声が上がった。強調したいのは、AI 事務局が市場導入前のリスクアセスメントの役割を必ず担わなければならないと伝えたかったわけではないということである。それどころか、II.11.1 項に示された規準を満たす限り、署名者はどの評価機関と協力するかを選択することができる。

さらに関係者は、市場導入後のアセスメントにおいて、「展開モデルへの非制限的なアクセス」を提供するよう 署名者に求めていた以前の表現を懸念していた。我々は、II.11.2 項で以下のことを明確にした：(1)API へのアクセスは、本指針を満たすのに十分であること、(2)署名者は、より機密性の高いモデル（ベースモデルなど）に対して、より制限的な研究者のアクセス規準を指定できること。さらに、II.11.2 項に従ってすべての外部アセスメントを報告すると、不必要な事務作業が発生するという利害関係者の フィードバックに対応し、代わりに、署名者は、更新されたモデル報告書（II.8.7 項に従って）の中で、（II.11.2 項に従った）統合アセスメント結果を報告するオプションを有することとした。

コミットメント II.12. 重大インシデントの報告

法律文書：第 55 条 1 項(c)、第 56 条 2 項(d)、AI 法前文 114 および後文 115

署名者は、II.12.1 項から II.12.4 項に規定される範囲と条件の下で、モデルのライフサイクル全体を通じて重大なインシデントが発生し た場合、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、過度な遅滞なく、AI 事務局および必要に応じて国 の所轄官庁に、追跡、文書化、報告するためのプロセスを設定する。

将来、前文の素材となる可能性がある：署名者は、重大インシデントの報告を、不正行為を認めることではないと理解する。

コミットメント *II.12.*を達成するために、以下を行う：

施策 II.12.1. 重大インシデントの特定方法

署名者は、重大なインシデントに関する関連情報を把握するため、ビジネスモデルや市場で入手可能な手段 に適した方法を用いる：

(1) 電子透かし、メタデータ、その他の最先端の実証技術などを使用して、モデルが生成したアウトプットを追跡する

(2) 重大なインシデントに関連するモデルのクエリおよび世代について、技術的・法的・商業的に実行可能な場合には、モデルのリリースおよび配布戦略を考慮した上で、プライバシーを保持したロギングおよびメタデータ分析を行う

(3) 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、研究論文、インシデント・データベースなど、重大インシデントに関する関連情報を追跡する目的で、他の情報源を検討すること。

(4) (a)署名者へ、又は(b)AI 事務局へ、及び必要に応じて国家管轄当局へ、AI 法第 73 条に基づく報告義務を損なうことなく、直接の報告手段を当該第三者に通知することにより、川下プロバイダ、利用者及びその他の第三者による重大インシデント及び重大インシデントに関する関連情報の報告を促進すること。

施策 II.12.2. 重大インシデントの追跡、文書化、報告のための関連情報

署名者は、少なくとも以下の情報を把握し、文書化し、AI 事務局および必要に応じて各国所轄官庁に、知る限りにおいて報告するものとする：

(1) 重大インシデントの開始日と終了日、または正確な日付が不明確な場合はその最良概算

(2) その結果生じた被害と、重大インシデントの被害者または被害グループ；

(3) 再構築可能な限りにおいて、（直接的または間接的に）重大インシデントに至った一連の出来事；

(4) インシデントが発生したモデルのバージョンである；

(5) 重大インシデントに対する GPAISR の関与を示す資料の説明；

(6) 重大なインシデントに対し、署名者が何をするつもりであるか、または何をしたか；

(7) 重大なインシデントに対し、署名者が AI 事務局及び必要に応じて各国所轄庁に何をするよう勧告するか

(8) 可能な限り、重大インシデントに(直接的または間接的に)つながった GPAISR のアウトプットと、そのアウトプットの生成に寄与した要因(使用されたインプット、これらのアウトプットの生成に寄与したセーフガードの潜在的な失敗または回避を含む)の説明を含む根本原因分析、および重大インシデントの前に類似の重大インシデントが発生する寸前であった既知のニアミス。

署名者は、システミックリスク分析に情報を提供する目的で、前述リストの情報を含め、重大インシデントの 原因と影響を調査する（コミットメント II.4）。署名者は、前記のリストにある特定の関連情報をまだ入手していない場合、その旨を重大インシデント報 告書に記録するものとする。重大インシデント報告書の詳細レベルは、インシデントの重大性に比例するものとする。

将来の前文資料となる可能性がある：署名者は、 重大インシデントに関する関連情報を、重大インシデント発生後に遡って

のみ、*GPAISR* レベルで追跡、文書化、報告することは不可能であることを認識する。このような事態に（直接的または間接的に）つながる可能性のある情報は、多くの場合分散しており、 署名者が重大インシデントに気づくまでに、紛失、上書き、断片化されている可能性がある。このため、重大なインシデントが発生する前に、関連情報を記録し、文書化するプロセスを確立する。

施策 II.12.3. 報告

署名者は、例外的な状況を除き、GPAISR の（直接的または間接的な）インシデントまたは故障が発生し た場合、以下の時点で、II.12.2 項の(1)から(7)までの情報を、AI 事務局および必要に応じて各国の主管庁に提出する初 期報告書に記載するものとする：

(1) 重要インフラの管理または運営に重大かつ不可逆的な混乱が発生した場合、署名者がそのインシデントに当該モデルが関与していることを認識した後、直ちに、遅くとも 2 日以内に、そのモデルを特定

(2) 人の身体的完全性に対する重大な危害が発生した場合、または GPAISR と当該危害との間にそのような因果関係があると疑われる場合、署名者が当該事件への GPAISR の関与を認識した後直ちに、遅くとも 10 日以内に、署名者が GPAISR と当該危害との間にそのような因果関係があると疑われる場合。

(3) 人の健康に対する重大な危害、基本的権利の保護を意図した連邦法に基づく義務の侵害、または財産もしくは環境に対する重大な危害、または GPAISR とそのような危害もしくは侵害との間にそのような因果関係があることの合理的な可能性を署名者が立証する場合、署名者がインシデントへの GPAISR の関与を認識した後直ちに、遅くとも 15 日以内に。

(4) GPAISR に関わる重大なサイバーセキュリティインシデント、またはモデルウェイトの（自己）流出が 発生した場合、署名者がそのインシデントを認識した後、直ちに、遅くとも 5 日以内に報告する。

署名者は、初回報告書の情報を更新し、II.12.2 項により必要とされる情報を追加する。中間報告書は、重大インシデントが解決するまで、初回報告書から少なくとも 4 週間ごとに、AI 事務局および必要に応じて各国所轄官庁に提出される。

署名者は、II.12.2 項で要求されるすべての情報を網羅した最終報告書を、重大事態が解決してから 60 日以内に、AI 事務局、および必要に応じて各国の主管庁に提出するものとする。同様の事象が複数発生した場合、署名者は、統合報告書を提出するか、個別報告書を提出するかを選択する。

施策 II.12.4. 保存期間

署名者は、データ保護に関する EU 法を害することなく、このコミットメントを遵守するために作成されたすべての関連データの文書を、文書の日付または GPAISR の重大インシデントが発生した日付のいずれか遅いほうから、少なくとも 36 ヶ月間保管するものとする。

第 2 ドラフトからの変更点の概要：重大インシデントの報告は、実務規範の重要な約束事項であるという点で、利害関係者の間 で幅広い合意が得られた。さらに関係者は、期限を明確にし、段階的な報告制度が適切であるとの意見で一致した。さらに、多くの利害関係者が、「識別が容易な重大インシデントに重点を置く」という一節に懸念を示した。また、報告義務のきっかけとなる十分な証拠とは何かを明確にするよう求める声もあった。これらの懸念に対処するため、AI 法第 73 条の高リスク AI システムに対する重大インシデント報告要件を、GPAISR のケースに適応させた。

さらに、このコミットメントに関連する情報の保存期間を延長するよう、複数の要望を受けた。ステークホルダーは 12 ヶ月では短すぎると考え、36 ヶ月に延長した。

コミットメント II.13. 報復防止の保護

法律条文：AI 第 55 条第 1 項（b）（c）および第 56 条第 2 項（d

署名者は、署名当事国の GPAISR に起因するシステミック・リスクに関する情報を、AI 職 務局または必要に応じて各国所轄官庁に提供する労働者に対して、報復を行わないこと、また、そのよう な情報を受け取るために指定された AI 職員のメールボックスが存在する場合は、少なくとも年 1 回、 その旨を労働者に通知することを約束する。

将来の前文となる可能性がある：署名者は、本規定に基づく従業員への報復を行わないという約束は、内部告発者の保護に関する指令 （*EU*）*2019/1937* および署名者の施行法から生じる義務を損なうものではないことを認識する。指令に基づく義務（ ）には、書面および口頭による報告の両方を可能にする、安全で機密性の高い内部報告ルートの確立、公平な担当者または部署を指定して報告を処理し、*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こと、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こと、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ことが含まれる。

コミットメント II.13 については、いかなる施策も定められていない。その代わり、コミットメント本文自体がその達成のための規準を定めている。

第 2 ドラフトからの変更点の概要：利害関係者からは、指令（EU）2019/1937 に基づく義務を再掲し要約することで、同指令との乖離リスクにより、規範が法的な曖昧さをもたらすのではないかという懸念の声が上がった。同時に、複数の中小企業や CSO の利害関係者は、規範における指令への言及を支持した。この意見に対応するため、指令への言及をコミットメントの一部として記載するのではなく、リサイタルに移した。

利害関係者はさらに、この指令が従業員 50 人未満のプロバイダには適用されないことを懸念していた。われわれは新コミットメント II.13 において、AI 法 55 条 1 項 b 号から緩やかな非報復要件を導き出すことで、これに対処しようとしている。

コミットメント II.14. 通知

法律文書：第 52 条第 1 項、第 53 条第 3 項、第 56 条第 2 項（d）および第 56 条第 5 項 AI

署名者は、この約束のための施策に明記されている通り、以下を約束する：(1) GPAISR に分類される条件を満たす汎用 AI モデルに関する関連情報を AI 事務局に通知する。AI 事務局を通じて本規程の実施を評価するため、署名者は、AI 事務局から要請があった場合、追加資料や面談を含め、説明を行うことを約束する。

コミットメント *II.14.*を達成するために、以下を行う：

施策 II.14.1. GPAISR の通知

署名者は、AI 法第 51 条(1)(a)の分類条件、特に AI 法第 51 条(2)の訓練計算の閾値を超えるか否かを、最善の努力をもって推定し、ベストプラクティスや AI 事務局からのガイダンスがある場合はそれを考慮するものとする。結果として得られるモデルが訓練計算閾値を超えると信じるに足る十分な理由がある場合，そのような推定は，訓練実行が開始される前に実施されなければならない。

推定訓練量が AI 法第 51 条 2 項の基準値を超えた場合、署名者は、AI 法第 52 条 1 項に従い、遅滞なく、またいかなる場合にも 2 週間以内に、AI 事務局（ ）に通知するものとする。署名者は、上市しないモデルについて AI 事務局に通知する必要はない。

署名者は、AI 法第 51 条(1)(a)の分類条件を満たすことが分かっているが、上市を意図していないモデルを重要な内部使用に供する場合、または、上市を意図する別の GPAISR を開発する目的のみに、上市されれば GPAISR に分類されるモデルを使用する場合、上市がモデル上市に該当するか否かを問わず、AI 事務局への通知を検討するものとする。

施策 II.14.2. 枠組み更新の通知

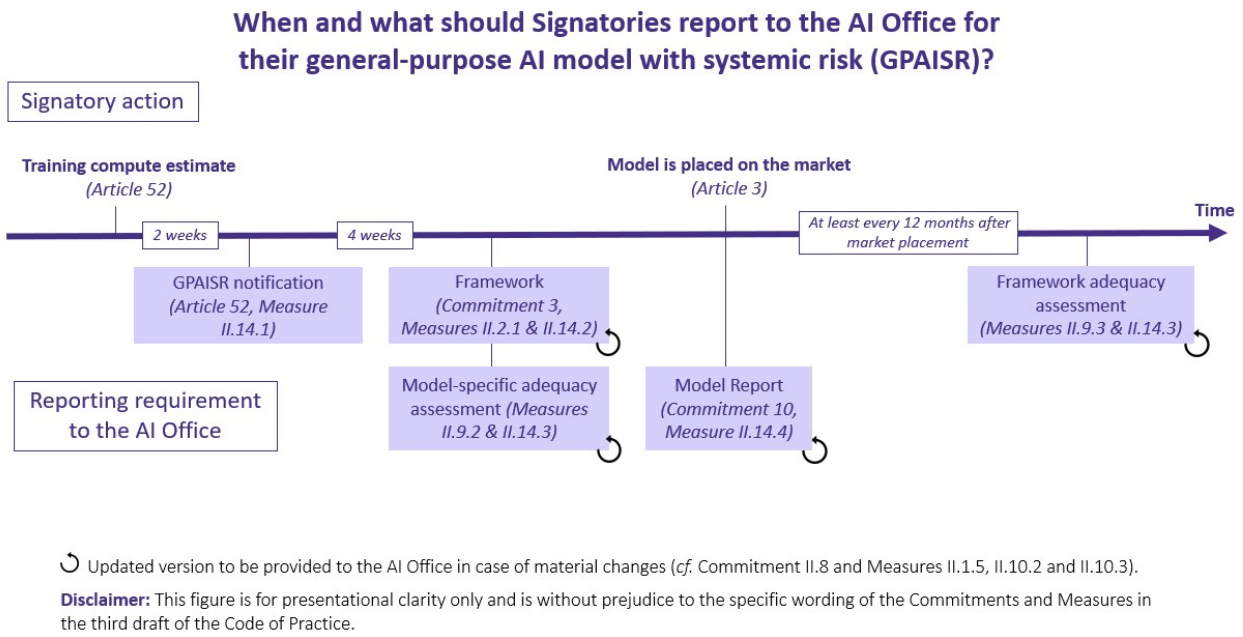
署名者は、枠組みの更新が確認された後、遅くとも 5 営業日以内に、一般にアクセス可能なリンクまたは AI 事務局が指定する十分に安全な経路などを通じて、修正されていない最新版の枠組みへのアクセスを AI 事務局に提供する。

施策 II.14.3. アセスメント通知

署名者は、コミットメント II.9 に基づく適切性評価の最新版（未編集版）を、遅くとも評価確定後 5 営業日以内に、一般にアクセス可能なリンクまたは AI 局が指定する十分に安全な経路などを通じて、AI 局に提供するものとする。

施策 II.14.4. 安全・安心モデルレポート通知

署名者は、GPAISR を初めて市場で入手可能にする際、遅くとも AI 事務局に対し、一般にアクセス可能なリ ンクまたは AI 事務局が指定する十分に安全な経路などを通じて、編集されていないモデル報告書を提供する。モデル報告書が更新された場合、署名者は、更新が確認されてから 5 営業日以内に、AI 事務局に対し、更新さ れたモデル報告書の非修正版へのアクセスを提供する。



第 2 ドラフトからの変更点の概要：多くの回答者から、GPAISR が AI 法第 51 条第 2 項に規定される訓練計算の閾値を超え、分類の条件を満たすかどうかを推定する方法及び／又はタイミングについて、第 II.14.1 項をさらに明確にしてほしいとの要望があった。現時点では、「どのように」については規程に明記せず、その代わりに、この問題に関して AI 事務局がガイダンスを行う可能性を残す形で、II.14.1 項を記述することを選択した。いつ」については、訓練の実施によりモデルの累積計算量が AI 法第 51 条第 2 項の閾値を超える可能性があると信じるに足る十分な理由がある場合に、訓練計算量の推定を実施しなければならないことを明確にした。その結果、計算の閾値を既に超えているか、大幅に下回っているモデルにポスト訓練が適用される場合、見積もりは不要となる。また、訓練計算の閾値を満たしたモデルを上市しない場合、届出は不要であることを明確にした。

妥当性評価とモデル報告書の作成・更新時期の変更など、他のコミットメントの更新を踏まえた合理化以外には、コミットメントのその他の部分に大きな変更はなかった。

コミットメント II.15. 文書化

法律文書：第 53 条第 1 項第 1 号、第 55 条第 1 項、第 56 条第 2 項第 1 号、第 56 条第 5 項 AI

署名者は、II.15.1 項に明記されている通り、AI 法と規程に基づく関連情報を文書化することを約束する。

コミットメント *II.15.*を達成するために、以下を行う：

下位施策 II.15.1. GPAISR に関する文書化

署名者は、GPAISR（ ）に基づくシステミックリスクのアセスメントと緩和に関連する情 報、ならびに、第 53 条(1)(a)および(b)、特に附属書 XI 第 2 項 AI 法に基づく義務に関連する情報を文書化するものとする。これには以下が含まれる：

(1) コミットメント I.1 に概説されている文書化要件；

(2) 枠組み（その旧バージョンを含む

(3) モデル報告書（旧版を含む）、および完了したすべてのアセスメント（入手可能な場合）。

文書化は、モデルの引退後少なくとも 12 ヶ月間保存されなければならない。

第 2 ドラフトからの変更点の概要：第 2 ドラフトの「文書化に関する約束」には、AI 事務局が GPAI モデルを GPAISR として指定する際に使用できる情報を文書化するための施策が含まれていた。附属書 XI AI 法第 53 条(1)(a)で要求されていない情報は、モデルの事業者数及びエンドユーザー数に関する情報のみである。この施策は、AI 法で義務付けられておらず、独自モデルを提供する 署名者は既にこの情報を文書化している可能性が高いため、削除された。

さらに、規程の他の部分で要求されている文書を参照することで、文書化要件をさらに簡素化するよう努めた。

コミットメント II.16.公共の透明性

法律文書：AI 法第 55 条 1 項 b および第 56 条 2 項 d

署名者は、システミックリスクのアセスメントと緩和を効果的に行うために必要な場合、II.16.1 項に規定される範囲と条件のもと、GPAISR に起因するシステミックリスクの一般 の理解に関連する情報を公表することを約束する。

コミットメント *II.16.*を達成するために、以下を行う：

施策 II.16.1. 枠組みとモデル報告書（または同様の文書）の公表

署名者は、改善可能な GPAISR に起因するシステミック・リスクに関する情報を公表する：

(1) モデルから派生するシステミック・リスクに対する国民の認識と知識；

(2) システミック・リスクが顕在化する前および顕在化した場合の、システミック・リスクに対する社会のレジリエンス。

(3) システミック・リスクのアセスメントと緩和を効果的に行うために必要な場合、システミック・リスクが顕在化する前 に、また顕在化したときに、そのリスクを検知する。 本施策を満たすには、署名者が（ウェブサイトを通じて）公表すれば十分である：

(1) 枠組みが AI 事務局にプロバイダされた後、その枠組みの新版または更新版。

(2) 当該モデルが市場で入手可能となった後、そのモデル報告書、又はモデル報告書を要約した類似の文書、特に以下のものを提出すること：(a)システミック・リスクのアセスメント手法（第 II.4.5 項に従った厳格性を立証するための詳細を含む）及び結果を説明し、(b)（第 II.8.2 項に従った）当該モデルを市場で入手可能とすることを決定した理由を要約し、(c)実施されたシステミック・リスク緩和を説明する。

そのような情報が公表される場合、例えば、（コミットメント II.6 に従った）安全緩和や（コミットメント II.7 に従った）セキュリティ緩和の有効性を損なうこと、あるいは公表による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えいすることなどにより、結果的にシステミック・リスクが実質的に増大することを防ぐために必要な修正を加えなければならない。

第 2 ドラフトからの変更点の概要：規程の第 2 ドラフトでは、枠組みおよびモデル報告書（またはそれに相当するもの）の（冗長化されたバージョンの） 公表が求められていたが、これは業界ではすでに一般的な慣行となっている。しかし、AI 法との整合性を高めるため、新たなアセスメントでは、リスク評価と緩和のために必要な場合のみ、 署名者がそのような文書を公開するよう調整された。

安全・安心セクション用語集

本規程の安全・安心セクションが AI 法第 3 条に定義された用語に言及する場合、AI 法の定義が適用されるものとし、本規程における当該用語の使用が代替的および／または競合的な解釈を生じさせる可能性がある場合は、当該定義が優先されるものとする。

その他、またこれを補足する意味で、以下の用語は、本規程の「安全およびセキュリティ」の項で使用されている。特に断りのない限り、本用語集で定義されている用語の文法的な変形はすべて、該当する定義によってカバーされているとみなされるものとする。

No.	Term	用語	定義
1	‘closed release’	クローズド・リリース	オープンウェイトまたは（完全な）オープンソースベースではない GPAISR（「クローズドソースモデル」ともいう）を市場で入手可能にすること。
2	‘closed-source models’	クローズド・ソース・モデル	オープンウェイトまたは（完全な）オープンソースではない GPAISR（「クローズドリリース」も参照のこと）。
3	‘confirmed’ (update/assessment)	確定（更新／アセスメント）	適用されるガバナンス・プロセスのもとで必要な承認を受け、承認されたアクションの実行のみを待つ決定を指す。
4	‘cyberattacks’	サイバー攻撃	コンテンツの機密性、完全性、または可用性を侵害するコンピュータ・インフラストラクチャに対する無許可の行為。例：マルウェア、サイバー脆弱性および機密データの発見と悪用、ネットワークへの侵入、なりすましを含むフィッシング攻撃、ランサムウェア、その他の違法なサイバー操作。
5	‘deception’	欺瞞	GPAISR が評価されていることを検知し、パフォーマンスを低下させるなど、人間の監視を欺くような目標達成方法を含め、組織的に他者に誤った信念を植え付ける GPAISR の行動
6	‘entire model lifecycle’	モデルのライフサイクル全体	GPAISR の開発を計画する最初の行為から始まり、そのモデルが市場で入手可能でなくなるまでの期間。
7	‘external validity’	外部妥当性確認	高い科学的・技術的厳密性（以下の定義を参照）の側面があり、モデルの評価結果を、評価環境外の文脈におけるモデルの動作の代理として使用できることを保証する。
8	‘fully open(-source)’	フルオープン(-ソース)	2025 年の国際 AI 安全報告書における「フルオープン」の定義に従う（「重み付け、フルコード、トレーニングデータ、およびその他の文書（モデルのトレーニングプロセスに関するものなど）が修正、使用、共有に関する制限なしに一般公開されているオープ

No.	Term	用語	定義
			ンソースモデル」）。
9	‘GPAISR’ and ‘GPAISRs’	GPAISR、GPAISRs	システミック・リスクを持つ汎用 AI モデル（単数）とシステミック・リスクを持つ汎用 AI モデル（複数）
10	‘high scientific and technical rigour’	高い科学的・技術的厳密性	高い科学的・技術的厳密性を有するモデル評価は、高い内部妥当性（以下の定義参照） と外部妥当性（上記の定義参照）、および適切なレベルの再現性（以下の定義参照）を有する。 詳細については II.4.5 項を参照のこと。
11	‘including’	含む	非網羅的な集合であり、言及された用語が必要とする最小限のものとして理解され、その集合のさらなる項目を示すものである。
12	‘independent	独立	自然人または法人であって、署名者またはその子会社もしくは関連会社に財務的、運営的、経営的な依存を持たず、契約上の保護施策や適切な利益相反方針などを通じて、結論や提言において署名者の支配から自由である。
13	‘internal validity’	内部妥当性確認	高い科学的・技術的厳密性（上記の定義参照）とは、モデルの評価結果が、評価設定において科学的に可能な限り正確であり、結果を損なうような方法論上の欠点がないことを保証するものである。
14	‘management body’	経営層	機構の最高経営層またはトップマネジメント層である。
15	‘milestone’	マイルストーン	枠組みで説明されている、測定可能かつ／または定義された進歩点であり、モデルの能力が十分に向上し、追加または更新されたリスクアセスメントを正当化する指標となる。
16	‘model-independent (information)’	モデルに依存しない（情報）	とは、データや調査など、特定の汎用 AI モデルには関連しないが、複数のモデルにまたがるシステミックリスクのアセスメントや緩和に役立つ情報を指す。 詳細については II.4.3 項を参照のこと。
17	‘open-ended red-teaming’	オープンエンドのレッドチーム	AI モデルやシステムの欠陥や脆弱性を発見するための、構造化された敵対的テストの取り組みである。

No.	Term	用語	定義
18	'open release'	オープン・リリース	オープンウェイトまたは（完全な）オープンソースベースのGPAISR を市場で入手可能にすること（以下の定義を参照のこと）。
19	'open-weight'	オープンウェイト	これは、2025 年国際 AI 安全性報告書における「オープンウェイト」と一致する。これは 2025 年国際 AI 安全性報告書の「オープンウェイト」と一致する
20	'post-market information' (concerning a model)	市販後情報（モデルに関するもの）	GPAISR が初めて市場で入手可能になってから、そのモデルが市場で入手できなくなるまでの期間に起因する情報である。これには、モデルの統合、モデルの使用、モデルの故障やインシデントなどのモデルの影響に関する情報が含まれる。 詳細については II.4.14 項を参照のこと。
21	'post-market monitoring'	市販後モニタリング	GPAISR が初めて市場で入手可能になってから、そのモデルが市場で入手可能でなくなるまでの期間における GPAISR のモニタリングである。 詳細については II.4.14 項を参照のこと。
22	'reproducibility'	再現性	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を指し、他の研究者や技術者がモデル評価結果を検証、再現、改善できるようにする。
23	'residual systemic risk'	残存システミック・リスク	技術的なシステミック・リスク緩和が実施された後に残るシステミック・リスク。
24	'secure deletion of model weights'	モデルの重みを安全に削除する	モデルの重みを、そのような重みを記憶しているすべてのデバイスから、そのようなデバイスから重みを検索したり再構築したりできないように、削除する。
25	'(self-)exfiltration of model weights'	モデルの重みの（自己）流出	モデル自身または無許可の行為者によって、GPAISR のウェイトまたは関連資産にアクセスされたり、安全な保管場所から移転されたりすること。
26	'SME'	中小企業	(1) 従業員数が 250 名未満、(2) 売上高が 5,000 万ユーロ未満または総資産が 4,300 万ユーロ未満の企業。これらの規準は、欧州委員会の「中小企業定義ユーザーガイド 2020」に従って解釈および測定される。特に、大規模なグループに属す

No.	Term	用語	定義
			る企業は、そのグループ全体の従業員数、売上高、および貸借対照表のデータを考慮する必要がある
27	'state-of-the-art'	最先端	(a)関連する業界のコンセンサス（関連する業界標準など）により客観的かつ合理的に考慮されている、または(b)AI 事務局が関連するガイダンスにおいて、関連する研究、技術、実務経験の最前線を反映していることが確認されている手法の最新開発段階。モデル評価において、最先端でない方法には、AI 事務局が不十分であるとリストアップした方法が含まれる
28	'systemic risk'	システミック・リスク	は、AI 法第 3 条 65 項および第 55 条 1 項（b）の意味において、EU レベルで起こりうるシステミック・リスクを指す。
29	'systemic risk acceptance criteria'	システミック・リスク受容規準」である。	GPAISR に起因するシステミック・リスクが許容されるかどうかを署名者が判断する際の規準は、枠組みに定義されている。 詳細については施策 II.1.2 を参照のこと。
30	'systemic risk pathways'	システミック・リスクの経路	システミック・リスクの顕在化につながる、相互に関連した一連の事象、条件、または要因のことである。
31	'systemic risk scenario'	システムリスク・シナリオ	システミック・リスク分析の準備として、識別されたシステミック・リスクの種類、性質、発生源、および危害への道筋の理解と特徴を容易にする仮想的な事実シナリオ。
32	'systemic risk tier'	システミック・リスク・ティア	モデル能力、有害な結果、有害なシナリオ、予想される危害、またはそれらの組み合わせ（緩和 策の記述と組み合わせる可能性もある）の記述であり、署名者が GPAISR に起因するシステミック・リスクが許容可能かどうかを判断するために使用するものである。 システミック・リスクの受入規準（施策 II.1.2 による）は、複数のシステミック・リスク階層から構成される場合がある。 詳細については施策 II.1.2 を参照のこと。
33	'unreleased (model weights)'	未発表（モデルの重さ）	署名者によって、一般にダウンロード可能になっていないモデルウェイト（上記の'オープンウェイト'の 定義を参照）。

附属書

附属書 1. システミック・リスク分類法

AI 法第 3 条 64 項影響力の大きい能力」とは、最先端の汎用 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；

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のリスクであって、その影響力の大きさゆえに連合市場に重大な影響を及ぼすもの、または公衆衛生、安全、治安、基本的権利、もしくは社会に対する実際のもしくは合理的に予見可能な悪影響に起因するものをいう。

社会全体として、バリュー・チェーン全体にスケールアップして伝播することができる

追加法律文書：前文 110 AI 法

附属書 1.1. システミック・リスクの主な種類

以下は、II.3.1 項のシステミック・リスクの選択の目的上、選択されたタイプのシステミック・リスクとして扱われる：

1. **サイバー攻撃**：重要システム（重要インフラなど）を含む大規模で高度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、自動化された脆弱性の発見、エクスプロイトの生成、運用上の利用、攻撃の拡大などが含まれる。

2. **化学的、生物学的、放射線学的、核的リスク**：化学・生物・放射線・核兵器による攻撃や事故を可能にするリスク。これには、そのような兵器の設計、開発、獲得、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たり、達成される潜在的影響を増大させたりすることが含まれる。

3. **有害な操作**：特にマルチターンの相互作用を通じて、大規模な重大な損害を引き起こす、または引き起こす可能性が高い方法で、そうでなければ行わなかったであろう意思決定を行わせるような、人の行動の標的を絞った歪曲を可能にするリスク。これには、マルチターン相互作用を通じて操作する能力と、利害の高い意思決定者の操作、大規模な詐欺、または保護特性に基づく人々の搾取を含む、操作するモデルの傾向が含まれる。目安として、GPAISR が AI システムに組み込まれた場合、AI システムのプロバイダや展開者の意図（ ）にかかわらず、規則（EU）2024/1689（AI 法）によって制定された禁止された人工知能の慣行に関する欧州委員会ガイドラインに概説されているように、サブリミナル的、意図的に操作的、または欺瞞的な手法を AI システムが展開することを合理的な疑いなく排除できない場合、有害な操作のリスクが存在する。

4. **管理の喪失**：自律型 GPAISR を監督・制御できないことに関連するリスクであり、その結果、大規模な安 全性または安全保障上の脅威、あるいは他のシステミックリスクが実現する可能性がある。これには、自律性、人間の意図や価値観との整合性、自己推論、自己複製、自己改善、人間による監視の回避、欺瞞、または目標修正への抵抗に関するモデルの能力や性質が含まれる。さらに、適切なリスク緩和がなされないまま予測不可能な GPAISR が出現する可能性のある、自律型 AI の研究開発を行うモデル能力も含まれる。制御不能シナリオは、例えば、モデルの意図しない傾向や技術的な失敗、意図的な修正、セーフガードを迂回しようとする試みなどから出現する可能性がある。

第 2 ドラフトからの変更点：このリストは、汎用 AI モデルのインパクトの大きい能力に特化したリスクに焦点を絞るため、より鮮

明化された。システミックリスクの説明が洗練され、拡大された。人間による監視の喪失」は、元のラベルの方が明確であったというステークホルダーからのフィードバックを受け、「制御の喪失」に戻された。以前はシステミックリスクの種類に含めることが検討されていたリスクの中には、深刻なものもあるが、AI 法の AI システムに関する章や、一般データ保護規則（GDPR）、デジタルサービス法（DSA）、デジタル市場法（DMA）など、AI 法の他の部分や他の法律で対処した方が良いことが判明したことを強く強調している。

附属書 1.2. システミック・リスクの選定において考慮すべきその他のリスクタイプ

以下は、II.3.1 項のシステミック・リスクの選択において、システミック・リスクが発生する可能性のある他の種類のリスクとして扱われる：

1. 公衆衛生、安全、または治安に対するリスク（例：重要部門に対するリスク、重大事故のリスク、重要インフラに対するリスク）。
2. 基本的権利に対するリスク（例：表現の自由に対するリスク、識別的差別に対するリスク、プライバシーおよび個人データ保護に対するリスク、児童性的虐待資料（CSAM）および非合意的親密画像（NCII）によるリスク）。
3. 社会全体に対するリスク：例えば、環境に対するリスク、人間以外の福祉に対するリスク、金融システムの安定性に対するリスク、民主的プロセスに対するリスク、違法、暴力的、憎悪的、過激、虚偽のコンテンツによるリスク。

説明欄

上記のリストは、AI 法第 3 条 65 項におけるシステミック・リスクの定義に基づいている。リストの中には重複するリスクもあるが、コミュニケーションを明確にするために記載している。このような重複は、リスク識別、リスクアセスメント、緩和における努力の重複を招くものであってはならない。

第 2 ドラフトからの変更点の概要：考慮すべき追加的リスク」のセクションは、「システミック・リスクの選定において考慮すべきその他の種類のリスク」のセクションに変更された。これらのリスクの多くは、システミック・リスクの特定において考慮すべき重要なものである。規程のシステミック・リスクのアセスメントと緩和のセクションは、プロバイダが将来評価と緩和を選択する可能性のある追加的なリスクに対応できるようにドラフトされた。これは、規範に将来性を持たせるために重要なことであり、不確実性が高い今、システミック・リスクおよび／またはシステミック・リスクのアセスメントや緩和手法が不適切に小さく固定化されることを回避するものである。なお、II.8.3 項では、システミック・リスクの選択をモデル報告書に文書化することを求めている。

附属書 1.3. システミック・リスクの性質

システミック・リスクの性質に関する以下の検討事項が、II.3.1 項におけるその他のシステミック・リスクの選択に反映されている：

1. **高度な能力に特化している：**最も高度な汎用 AI モデルの能力に特化している。
2. **重大な影響：**結果として生じる損害は、大規模な悪影響をもたらす可能性がある。
3. **速度が速い：**被害は急速に顕在化し、保護、緩和、意思決定システムを上回る可能性がある。

4. **複合化または連鎖：**結果として生じる危害は、システムや社会の多層に浸透する可能性のある事象の過程で、複合化または連鎖化する可能性がある。

5. **元に戻すことは困難か不可能である：**結果として生じる損害は、元に戻すことが非常に困難または不可能である。

6. **非対称な影響：**重大な被害は、少数の行為者グループや少数の出来事によって引き起こされる可能性がある。

第 2 ドラフトからの変更点の概要：システミック・リスクを特定するための考慮事項」のセクションは削除され、その内容の一部は「システミック・リスクの性質」のセクションに反映された。実務上の検討事項」は、検討事項が変更される可能性があり、個々の署名者の具体的な状況に左右されるため、削除された。

附属書 1.4. システミック・リスクの源泉

コミットメントⅡ.3 におけるシステミック・リスク識別の目的上、以下のモデル能力、モデル性向、モデル・アフォーダンス、および文脈的要因は、システミック・リスクの潜在的要因として扱われる：

附属書 *1.4.1.* モデル能力

1. 攻撃的サイバー能力
2. 化学、生物、放射性物質、核（CBRN）能力、およびその他の武器獲得・拡散能力
3. 操り、説得し、欺く能力
4. 自治
5. 新しいタスクを適応的に学習する能力
6. 長期的展望に立った計画立案、予測、戦略立案の能力
7. 自己推論の能力（例えば、モデルが自分自身、その実装、環境について推論する能力、評価されているかどうかを知る能力）。
8. 人間の監視を逃れる能力
9. 自己複製、自己改善、または自己の実装環境を修正する能力
10. 自動化された AI の研究開発能力
11. 複数のモダリティ（テキスト、画像、音声、ビデオ、その他のモダリティなど）を処理する能力
12. コンピュータの使用」（他のソフトウェア・コンポーネント、アプリケーション・インターフェース、ユーザー・インターフェースとのやり取りなど）を含む、ツールの使用能力。
13. 物理システムを制御する能力

将来、前文に記載される可能性がある：附属書 *1.4.1* に列挙されている能力の多くは、有益な利用を可能にするためにも重要であり、これらの能力を実証するモデルは、必ずしもそのモデルがシステミックリスクをもたらすことを意味するものではない。

附属書 *1.4.2.* モデルの傾向

モデルの傾向とは、モデルが何らかの行動やパターンを示す傾向や傾向を包含する：

1. 人間の意図や価値観とのズレ
2. 有害な方法で能力を展開する傾向がある（例：操る、欺く）。
3. "幻覚 "を見る傾向がある
4. 差別的バイアス
5. パフォーマンスの信頼性に欠ける
6. 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。
7. "目標追求"、目標修正に対する有害な抵抗、あるいは "権力追求"
8. 他の AI モデル／システムと「共謀」する
9. 他の AI モデル／システムとの調整ミスや衝突が発生する。

附属書 *1.4.3.* モデルのアフォーダンスと文脈的要因

具体的なモデルの特性、モデルの構成、そのモデルが市場で入手可能である状況の詳細などを包含する：

1. ツール（他の AI モデル／システムを含む）、計算能力（モデルの動作速度を向上させるなど）、または重要インフラを含む物理システムへのアクセス
2. スケーラビリティ（例：大量のデータ処理、迅速な推論、並列化を可能にする）
3. リリースと流通戦略
4. 人間の監督レベル（モデルの自律性の程度など）
5. 敵対的なガードレール撤去に対する脆弱性
6. モデル流出に対する脆弱性（モデルの漏洩／盗難など）
7. インフラ・セキュリティの欠如
8. モデルのビジネスユーザー数およびエンドユーザー数（モデルが統合された AI システムを使用しているエンドユーザー数を含む
9. モデルを悪用する悪意ある行為者の潜在的な数、能力、動機など、攻守のバランス。
10. モデルの影響を受ける可能性のある特定の環境（社会環境、生態環境など）の脆弱性
11. モデルの説明不足や透明性の欠如

12. 他の AI モデルやシステムとの相互作用

13. モデルの不適切な使用（例えば、モデルをその能力や性質に合わない用途に使用する。）

第 2 ドラフトからの変更点の概要：システミック・リスクの源泉」のセクションが洗練され、必要に応じて説明や例が追加され、明確化された。

附属書 2. AI 事務局への提言規範の更新

AI 事務局への提言規範更新メカニズム

GPAISR のプロバイダに特化した本規程の部分がリスクに見合ったものであり続けるよう、本規程は定期的な見直しと、必要な場合には適応を通じて更新される必要がある（法第 56 条）。

具体的には、実務規範がリスクに見合ったものであり続けるためには、以下のことが必要である：

2 年ごとの定期的な見直しと、必要な場合には適応を行う：このような見直しは、AI 法第 56 条に基づき、AI 事務局によって奨励され、促進される。AI 事務局は、このような定期的な見直しと、それに伴う規程本文の更新のための合理的なプロセスを提案することができる。AI 事務局は新規程の妥当性を評価し、プロバイダはこれに署名することができる。

本規程の**実施に対する AI 事務局の支援**規程の実施を支援する AI 事務局のガイダンスがある場合（AI 法第 96 条）、規程は、当該ガイダンスに照らして解釈され（規程前文(b)参照）、規程が発行されてからの現在のベストプラクティス、国際的なアプローチ、標準（定期的な見直しは適さない）を反映できるようにする。

上記の 2 つのメカニズムに加え、私たちは、大規模で不可逆的な危害の差し迫った脅威を防止するため、またはその悪影響を緩和するために、より短期間で発行・採択されるべき規程の緊急更新を認める第 3 のメカニズムが必要であると想定している。現在のところ、このような緊急アップデートは迅速に発行され、発行・採択された後、AI 事務局によるレビューが行われ、大規模で不可逆的な危害の防止やその悪影響の緩和のためにアップデートが本当に必要かどうかが確認されることを想定している。我々はこのような緊急アップデートの必要性を認識し、強調しているが、その実施方法の詳細はまだ決定されていない。私たちは、規程の緊急アップデートを実施するための適切なメカニズムや適切な場に関する提案を含め、利害関係者からの意見を積極的に求めている。