Check for updates

NIST 特別刊行物 800
NIST SP 800-218A

生成的 AI とデュアルユース基盤モデルのための安全なソフトウェア開発プラクティス

SSDF コミュニティプロフィール

ハロルド・ブース
ムルギア・スッパヤ
アポストル・ヴァシレフ
マイケル・オガタ
マーティン・スタンレー
カレン・スカーフォン

本書は以下から無料で入手できる：
https://doi.org/10.6028/NIST.SP.800-218A

NIST 特別刊行物 800
NIST SP 800-218A

生成的 AI とデュアルユース基盤モデルのための安全なソフトウェア開発プラクティス

SSDF コミュニティプロフィール

ハロルド・ブース
ムルギア・スッパヤ
アポストル・ヴァシレフ
コンピュータ・セキュリティ・ディビジョン
情報技術研究所

マーティン・スタンレー
サイバーセキュリティとインフラセキュリティ
エージェンシー（*CISA*）

マイケル・オガタ
応用サイバーセキュリティ部門 情報技術研究所

カレン・スカーフォン
スカーフォン・サイバーセキュリティ

本書は以下から無料で入手できる：
https://doi.org/10.6028/NIST.SP.800-218A
2024 年 7 月

米国商務省 ジーナ・*M*・ライモンド長官
国立標準技術研究所理事兼標準技術担当商務次官 ローリー・*E*・ロカシオ

本書では、実験手順を適切に特定するために、商用、非商用を問わず、特定の機器、装置、ソフトウェア、または材料を識別している。このような識別は、NIST によるいかなる製品またはサービスの推奨または保証を意味するものではなく、また識別された材料または装置が必ずしもその目的に対して利用可能な最良のものであることを意味するものでもない。

本書には、NIST がその法定責任に従って現在開発中の他の出版物を参照する場合がある。概念や方法論を含む本書の情報は、そのような関連出版物の完成前であっても、連邦機関によって使用される可能性がある。従って、各出版物が完成するまでは、現行の要求事項、ガイドライン、手順が存在する場合は、それらが引き続き有効である。計画と移行を目的として、連邦機関は NIST によるこれらの新出版物の開発を注意深く見守ることが望ましい。

各組織は、パブリックコメント期間中にすべての出版物のドラフトを検討し、NIST にフィードバックを提供することが推奨される。上記以外の多くの NIST サイバーセキュリティ出版物は、https://csrc.nist.gov/publications で入手できる。

認可

本書は、2014 年連邦情報セキュリティ近代化法（FISMA）（44 U.S.C. § 3551 et seq.、公法（P.L.）113-283）に基づく NIST の法的責任に基づき、NIST が作成したものである。NIST は、連邦情報システムの最低要件を含む情報セキュリティ標準およびガイドラインを策定する責任を負うが、そのような標準およびガイドラインは、そのようなシステムに対する政策権限を行使する適切な連邦当局者の明示的な認可がない限り、国家安全保障システムには適用されない。本ガイドラインは、行政管理予算局（OMB）サーキュラーA-130 の要件と一致している。

本書のいかなる内容も、法的認可に基づき商務長官が連邦政府機関に義務付け、拘束力を持たせた標準やガイドラインと矛盾するものであってはならない。また、本ガイドラインは、商務長官、OMB 長官、その他の連邦政府当局者の既存の権限を変更したり、これに取って代わるものと解釈されるべきではない。 本書は、非政府組織が任意で使用することができ、米国における著作権の対象ではない。ただし、NIST はその帰属を認める。

NIST テクニカルシリーズ方針

著作権、使用およびライセンスに関する声明

NIST 技術シリーズ出版識別子の構文

出版の歴史

2024-07-25、NIST 編集審査委員会により承認された。

この NIST テクニカルシリーズ出版物の引用方法

Booth H, Souppaya M, Vassilev A, Ogata M, Stanley M, Scarfone K (2024) Generative AI と Dual-Use Foundation AI モデルのための安全な開発プラクティス：An SSDF Community Profile.(National Institute of

Standards and Technology, Gaithersburg, MD), NIST 特別刊行物 (SP) NIST SP 800-218A. https://doi.org/10.6028/NIST.SP.800-218A.

認可 ORCID iDs

ハロルド・ブース 0000-0003-0373-6219

Murugiah Souppaya: 0000-0002-8055-8527 Apostol Vassilev: 0000-0002-9081-3042 Michael Ogata: 0000-0002-8457-2430

カレン・スカーフォン 0000-0001-6334-9486

連絡先

ssdf@nist.gov

国立標準技術研究所
担当：応用サイバーセキュリティ部門、情報技術研究所
100 Bureau Drive (Mail Stop 2000) Gaithersburg, MD 20899-2000

追加情報

関連コンテンツ、更新の可能性、文書の履歴など、本書に関する追加情報は https://csrc.nist.gov/pubs/sp/800/218/a/final。

すべてのコメントは、情報公開法（FOIA）に基づき公開される。

要旨

本文書は、セキュアソフトウェア開発フレームワーク（SSDF）バージョン 1.1 で定義されたセキュアソフトウェア開発プラクティスとタスクを補強するものであり、ソフトウェア開発ライフサイクル全体を通じて AI モデル開発に特化したプラクティス、タスク、推奨事項、考慮事項、注記、および参考文献を追加している。これらの追加事項は、大統領令（EO）14110「人工知能の安全、セキュア、かつ信頼できる開発と使用」をサポートするために、SSDF コミュニティ・プロファイルの形で文書化されている。このコミュニティ・プロファイルは、NIST に「生成的 AI とデュアルユース基盤モデルのための安全な開発プラクティスを組み込んだ（SSDF の）付属リソースを開発する」ことを課している。このコミュニティ・プロファイルは、AI モデルの製作者、それらのモデルを使用する AI システムの製作者、およびそれらの AI システムの取得者にとって有用であることを意図している。このプロファイルは、NIST 特別刊行物（SP）800-218「*Secure Software Development Framework (SSDF) Version 1.1*」と併せて使用されるべきである：ソフトウェアの脆弱性リスクを低減するための推奨事項」と併せて使用すること。

キーワード

人工知能、人工知能モデル、サイバーセキュリティリスクマネジメント、生成的人工知能、安全なソフトウェア開発、安全なソフトウェア開発フレームワーク（SSDF）、ソフトウェア取得、ソフトウェア開発、ソフトウェアセキュリティ。

コンピュータ・システム技術に関するレポート

国立標準技術研究所(NIST)の情報技術研究所(ITL)は、国家の測定および標準機構に技術的リーダーシップを提供することにより、米国経済と公共の福祉を促進する。ITL は、情報技術の開発と生産的使用を促進するために、試験、試験方法、参照データ、概念実証実装、技術分析を開発している。ITL の責任には、連邦情報システムにおける国家安全保障関連情報以外のコスト効率の良いセキュリティとプライバシーのための管理、管理、技術、物理的標準とガイドラインの開発が含まれる。Special Publication 800 シリーズは、情報システム・セキュリティに関する ITL の研究、ガイドライン、アウトリーチ活動、および産業界、政府、学術団体との協力活動について報告している。

想定読者

この文書の主な読者は 3 つある：

- ***AI* モデル製作者** - EO 14110 で定義されているように、独自の生成的 AI およびデュアルユース基盤モデルを開発している組織。
- ***AI* システムプロデューサー** - 生成的 AI またはデュアルユース基盤モデルを活用したソフトウェアを開発している組織
- *AI* システム取得者[1] - 1 つまたは複数の AI システムを利用する製品またはサービスを取得する組織

AI モデルのための安全なソフトウェア開発プラクティスについて理解を深めたいと考えている人々にとっても、この文書は有益であろう。

読者は、安全なソフトウェア開発や AI モデル開発の専門家であることは期待されていないが、これらの推奨される実践方法を実施するためには、そのような専門知識が必要になるかもしれない。

読者の皆様へ

標準開発組織（SDO）や、AI モデル開発のためのセキュアプラクティス一式を定義している他の組織の方で、その標準やガイダンスを SSDF プロファイルにマッピングしたい場合は、認可担当者（ssdf@nist.gov）までご連絡いただきたい。このプログラムでは、既存の参考文献を補強するためのマッピングを提出することができる。

また認可者は、この文書のどの部分についても、またこの文書に加えるべき実施例や参考文献の提案についても、いつでもフィードバックを歓迎する。すべてのフィードバックは、ssdf@nist.gov。

商標情報

[1] 生産者（producer）」と「取得者（acquirer）」という用語は、NIST SP 800-218 の「聴衆（Audience）」の記述との整合性のために選択された。

すべての登録商標はそれぞれの団体に帰属する。

謝辞

本書の作成にあたり、多数のパブリックコメントやその他の示唆に富む意見を寄せていただいたすべての組織および個人の方々に、認可に感謝する。大統領令（EO）14110「人工知能の安全、セキュア、かつ信頼できる開発と使用」に対応して、NIST は 2024 年 1 月にワークショップを開催し、講演者と出席者は、AI モデル開発とそれを活用するソフトウェアのユニークな側面に対応するために、セキュアなソフトウェア開発のプラクティスとタスクを適合させるための提案を共有した。また、本書の策定に資する提案やフィードバックを提供してくれた NIST の同僚や外部の専門家にも感謝する。

特許開示通知

注意：ITL は、本出版物のガイダンスまたは要件に準拠するために使用が必要となる特許請求の範囲の保有者に対し、当該特許請求の範囲を ITL に開示するよう要請している。ただし、特許権者は ITL の特許募集に応じる義務はなく、ITL は、本出版物に該当する特許があるとしても、それを特定するための特許調査を行っていない。

本書の発行日および、本書のガイダンスまたは要件に準拠するために使用が必要とされる可能性のある特許請求の範囲の識別を求めた後の時点で、ITL にそのような特許請求の範囲は識別されていない。

ITL は、本出版物の使用において特許侵害を回避するためのライセンスが不要であることを表明するものではなく、また示唆するものでもない。

目次

1.序文 9

1.1 目的 9

1.2 適用範囲 10

1.3 専門知識の源泉 10

1.4 文書の構造 11

2. SSDF コミュニティ・プロフィールを使う 12

3. AI モデル開発のための SSDF コミュニティ・プロフィール 15

附属書 A. 用語集 35

表のリスト

表 1. AI モデル開発のための SSDF コミュニティ・プロフィール

1.序文

大統領令（EO）14110 の 4.1.a 項「人工知能の安全、セキュア、かつ信頼できる開発および使用」[1]は、NIST に「生成的 AI およびデュアルユース基盤モデルのための安全な開発プラクティスを組み込むために、セキュアソフトウェア開発フレームワークのコンパニオンリソースを開発する」ことを課している。本文書はその付属リソースである。

AI モデルや AI システムのソフトウェア開発と使用は、他のデジタルシステムと同じリスクを多く受け継いでいる。このコミュニティにとってユニークな課題は、システムコードとシステムデータとの間の従来の境界が曖昧であること、またシステムとのインタラクションの手段として平易な人間の言語を使用することである。AI モデルやシステム、その設定パラメータ（モデルの重みなど）、そしてそれらが相互作用するデータ（学習データ、ユーザークエリなど）は、閉じたループを形成し、意図しない機能のために操作される可能性がある。

AI のモデルやシステム開発は、厳密な科学というよりはまだ芸術の域を出ておらず、開発者はモデルコード、トレーニングデータ、その他のパラメータを何度も繰り返し操作する必要がある。トレーニングデータセットは、未知の信頼できないソースから取得される可能性がある。モデルの重みとその他のトレーニング・パラメータは、悪意のある改ざんの影響を受けやすい。モデルによっては、複雑なため容易に徹底的な検査を行うことができず、検知されずに任意のコードが実行される可能性がある。ユーザークエリは、望ましくない、あるいは好ましくない出力を生成するように細工される可能性があり、適切にサニタイズされなければ、インジェクションスタイルの攻撃に活用される可能性がある。この文書の目的は、これらの新しいリスクに対処するために必要なプラクティスとタスクを特定することである。

1.1 目的

SSDF は、ソフトウェア開発ライフサイクル全体を通して、安全なソフトウェア開発のプラクティスを記述するための共通言語を提供する。この文書は、生成的 AI とデュアルユース基盤モデル開発に特化した推奨事項、考慮事項、注釈、参考文献を追加することにより、SSDF バージョン 1.1 で定義された実践とタスクを補強するものである。これらの追加は、*SSDF* コミュニティプロファイル（"プロファイル"）の形で文書化されている。これは、特定のユースケースに対応するために拡張された SSDF のプラクティスとタスクのベースラインである。例えば、「安全なコードストレージには、AI モデル、モデル重み、パイプライン、報酬モデル、および機密性、完全性、可用性の保護が必要なその他の AI モデル要素を含めるべきである。

このプロファイルは，SSDF version 1.1 がすでに含んでいるものを補足するものである。このプロファイルは NIST 特別刊行物 (SP) 800-218, *Secure Software Development Framework (SSDF) Version 1.1*：ソフトウェアの脆弱性リスクを低減するための勧告 [6]であり、 SP 800-218 なしで使用してはならない。また、一般的に AI モデルおよび AI システムの開発にも適用できるため、読者は、あらゆる種類のソフトウ ェア開発における SSDF の各実践方法およびタスクの実行方法に関する追加情報については、SP 800-218 で定義されている実装例および参考文献を活用する必要がある。

1.2 適用範囲

このプロファイルの適用範囲は、*AI* モデルの開発であり、これには、AI モデルのデータ調達、設計、訓練、微調整、評価、および他のソフトウェアへの AI モデルの組み込みと統合が含まれる。SSDF バージョン 1.1 および EO14110 と一致して、**AI モデルを用いた AI システムの配備と運用に関するプラクティスは対象外である**。同様に、AI モデルの開発に使用されるトレーニングデータやその他のデータに対するサイバーセキュリティの実践は対象であるが、その他のデータガバナンスと管理のライフサイクルは対象外である。

なぜなら、すべてのソースコードは使用前に脆弱性やその他の問題について評価されるべきであると想定されているからである。

1.3 専門知識の源泉

この文書は、以下のような数多くの専門知識を活用し、統合したものである：

- 人工知能リスクマネジメントフレームワーク（*AI RMF 1.0*）*[2]*、敵対的機械学習*[3]*、人工知能におけるバイアスの特定と管理のための標準に向けて*[4]*、*Dioptra* 実験テストベッド*[5]*など、信頼できる責任ある *AI* に関する *NIST* の研究および出版物： *Dioptra experimentation testbed for security evaluations of machine learning algorithms [5]* などがある。

- NIST のセキュアソフトウェア開発フレームワーク（*SSDF*）バージョン *1.1* [6]は、基本的で健全かつセキュアなソフトウェア開発のプラクティスをまとめたものである。これは、基本的かつ健全でセキュアなソフトウェア開発手法のセットであり、ソフトウエア製造者やソフトウエア取得者を含む利害関係者間のコミュニケーショ ンを円滑にするための共通言語を提供するものである。また、SSDF は、ソフトウェア・サプライ・チェーンのセキュリテ

ィを強化するために、EO 14028「*Improving the Nation's Cybersecurity*」[7] の支援にも使用されている。

- NIST サイバーセキュリティフレームワーク*(CSF)2.0* [8]、情報システムと組織のためのセキュリティとプライバシー統制 [9]、システムと組織のためのサイバーセキュリティサプライチェーンリスクマネジメントの実践 [10]など、NIST のサイバーセキュリティ全般に関するリソース。

- AI モデル開発者、AI 研究者、AI システム開発者、産業界および政府のセキュアソフトウェア実務者は、AI モデル特有のセキュリティ上の課題と、その課題に対処するためのプラクティスに関する専門知識を有している。この専門知識は、主に NIST が 2024 年 1 月に開催したワークショップを通じて得られたものであり、講演者と出席者は、AI モデル開発とそれを活用するソフトウェアに特有の側面に対応するために、セキュアソフトウェア開発のプラクティスとタスクを適合させるための提案を共有した。

1.4 文書の構造

この文書の構成は以下の通りである：

- セクション 2 では、SSDF の背景を説明し、SSDF コミュニティ・プロフィールとは何か、どのように利用できるかを説明する。

- セクション 3 は、AI モデル開発のための SSDF コミュニティ・プロファイルを定義している。

- 参考文献のセクションには、本文書で引用したすべての参考文献が記載されている。

- 附属書 A は、本文書で使用されている用語の解説である。

2. SSDF コミュニティ・プロフィールを使う

AI モデル製作者、AI システム製作者、AI システム取得者、その他は、ソフトウェア開発ライフサイクル全体を通じて、安全な AI モデル開発に関するコミュニケーションを促進するために SSDF を利用することができる。[2]SSDF のプラクティスに従うことは、AI モデル製作者が AI モデル内の脆弱性の数を減らし、未検出または未対処の脆弱性が悪用されることによる潜在的な影響を減らし、再発を防止するために脆弱性の根本原因に対処するのに役立つはずである。AI システム製作者は、AI モデル製作者とコミュニケーションする際に、AI モデル開発におけるセキュリティ対策について、また、開発中のソフトウェアに AI モデルを統合する際に、SSDF の共通語彙を使用することができる。また、AI システムの買収者は、買収プロセスなどにおいて、自社のサイバーセキュリティ要件やニーズを AI モデル製作者や AI システム製作者にうまく伝えるために SSDF の用語を使用することができる。

SSDF コミュニティプロファイルは、従うべきチェックリストではなく、AI モデルを含む安全なソフトウェア開発手法を採用するためのリスクベースのアプローチを計画し、実施するための出発点である。すべての実践とタスクがすべてのユースケースに適用できるわけではないので、プロファイルの内容は適応され、カスタマイズされることを意図している。組織は、AI モデル製作者、AI システム製作者、AI システム取得者としての組織の観点から、ソフトウェア開発プラクティスの脅威を軽減するために、どのようなプラクティスとタスクが関連し、適切で、効果的であるかを判断するために、リスクベースのアプローチを採用すべきである。リスク、コスト、実現可能性、適用可能性などの要素を考慮して、どのプラクティスとタスクを使用し、それぞれにどれだけの時間とリソースを割くかを決定する必要がある。AI モデル開発に固有のコストを効果的に考慮するために、コストモデルを更新する必要があるかもしれない。セキュアなソフトウェア開発に対するリスクベースのアプローチは、組織が AI モデルやシステムに関連する新たな機能やリスクに対応するにつれて、時間の経過とともに変化する可能性がある。

生成的 AI とデュアルユースの基盤モデルは、モデルのバージョニングとリネージを追跡する上で、さらなる課題を提示する。モデル・アーキテクチャを定義し、モデル・バイナリを構築するためのソース・コードは、バージョン管理、リネージ、再現性のための安全なソフトウェア・エンジニアリングのプラクティスに従うことができる。しかし、最終的なモデルの重みは、モデルが訓練

[2] SSDF 1.1 との整合性のため、本文書では一般的なソフトウェア開発ライフサイクルを使用する。この文書を使用する組織は、使用している機械学習固有のライフサイクルに適合させることが推奨される。

され、微調整された後にのみ定義される。このため、収集、処理、訓練のすべての側面を追跡することには限界がある。組織は、モデルの完全にカバーできる部分については安全なソフトウエア開発プラクティスに従 うべきであり、そのようなセキュリティ保証を得ることが困難な段階と対応する成果物については、 可能な限り安全なプラクティスを導入するよう努めるべきである。組織は、安全なソフトウエア開発プラクティスが適用されない部分と成果物を文書化する。

このプロファイルのプラクティス、タスク、推奨事項、および考慮事項は、継続的インテグレーション／継続的デリバリー（CI/CD）パイプライン内の他のソフトウェア資産とともに、機械学習オペレーション（MLOps）に統合することができる。

プロファイルにおける SSDF 実践の責任は、複数の組織間で共有される可能性がある。例えば、ある AI モデルがある組織によって作成され、第二の組織がホストする AI システムで実行され、それが他の組織によって使用される。

このような状況では、AI モデル製作者、AI システム製作者、AI システム取得者が関与する責任共有モデルが存在する可能性が高い。AI システム取得者は、AI システム・プロデューサーおよび／または AI モデル・プロデューサーとの間で、どの当事者が各業務やタスクに責任を持ち、各当事者がどのように契約への適合を証明するかを明記した契約を結ぶことができる。

SSDF と本プロファイルの限界は、サイバーセキュリティリスクマネジメントにしか対応していないことである。AI システムには、成熟したエンタープライズ・リスク・マネジメント・プログラムの一環として、組織がサイバーセキュリティリスクとともに管理すべき他の種類のリスク（データ・プライバシー、知的財産、バイアスなど）が数多く存在する。他の種類のリスクの識別とマネジメントに関する NIST のリソースには、以下のものがある：

- *AI* リスクマネジメントフレームワーク（*AI RMF*）[2]と *AI* リスクマネジメントフレームワーク プレイブック [11]がある。
- 敵対的機械学習：攻撃と低減の分類と用語 [3]
- 人工知能リスクマネジメントフレームワーク：生成的人工知能プロファイル[12]。
- 人工知能におけるバイアスを識別・管理するための標準に向けて [4]
- システムと組織のためのサイバーセキュリティサプライチェーンリスクマネジメントの実践 [10]。

- *NIST* プライバシーフレームワーク：*A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0* [13]（エンタープライズリスクマネジメントを通じてプライバシーを改善するためのツール）。

- サイバーセキュリティとエンタープライズリスクマネジメント（*ERM*）の統合 [14] [14

3. AI モデル開発のための SSDF コミュニティ・プロフィール

表 1 、AI モデル開発のための SSDF コミュニティ・プロファイルを定義する。各列の意味は以下の通りである：

- プラクティスには、プラクティスの名前と固有の識別子が含まれ、その後にプラクティスとは何か、なぜそれが有益なのかについての簡単な説明が続く。

  タスクは、練習を実行するために必要な 1 つ以上のアクションを指定する。各タスクには一意の識別子と簡単な説明が含まれる。

  SSDF1.1 からの変更」または「SSDF1.1 の一部ではない」というタグが明示されていない限り、すべての実施項目とタスクは SSDF1.1 から変更されていない。例えば、PW.3 の「使用前のトレーニング、テスト、微調整、およびデータの整合性を確認する」というプラクティスと、そのすべてのタスクである。

- **優先順位は、プロファイルの**文脈における各タスクの相対的な重要性を反映したものであり、各組織が独自の優先順位を割り当てるための出発点となることを意図している：

  - **高：**AI モデル開発の安全性は、他のタスクと比較して決定的に重要である。
  - **中** AI モデル開発の安全性を直接サポートする
  - **低：**安全なソフトウェア開発にとって有益であるが、一般的に他の多くのタスクよりも重要でない。

- **AI モデル開発に特有の推奨事項、考慮事項、および注意**事項には、特定のタスクについて、何を行うべきかを推奨したり、追加的な考慮事項を記載したりする 1 つまたは複数の項目が含まれる場合がある。組織は、セクション 2 に記載されたリスクベースのアプローチの一環として、必要に応じて項目を適応、カスタマイズ、省略することが期待される。

  各アイテムは、以下のいずれかで始まる ID を持っている：

  - 「R」（推奨：組織が行うべきこと）
  - 「C」（検討：組織が行うことを検討すべきこと）
  - 「N」（注：推奨事項と考慮事項以外の追加情報）

R、C、または N の指定とその番号をタスク ID に付加して、一意な識別子を作成することができる（例えば、「PO.1.2.R1」はタスク PO.1.2 の最初の勧告である）。

この欄の値が「SSDF 1.1 への追加なし」である場合は、このプロファイルにはそのタスクの AI モデル開発に固有の推奨事項、考慮事項、または注意事項が含まれていないことを示す。問題の安全な開発タスクに関する基本ガイダンスについては SSDF version 1.1 [6] を、タスクに関連する追加情報についてはこの文書の他の参考文献を参照すること。

- **参考文献（Informative References）は、**標準、ガイダンス、その他のコンテンツの一部を指し示す（マップする）ものであり、特定のタスクを実行する際の要求事項、推奨事項、考慮事項、その他の支援情報を含んでいる。参考文献の出典は以下の通りである：
    - *AI* リスクマネジメントフレームワーク *1.0* [2]がある。AI RMF と他のガイダンスや標準との間には、すでにいくつかのクロスウォークが定義されている。現在のセットについては、https://airc.nist.gov/AI_RMF_Knowledge_Base/Crosswalks を参照のこと。
    - *OWASP: Top 10 for LLM Applications Version 1.1* [15]。各識別子は、上位 10 脆弱性タイプの 1 つを示し、その脆弱性タイプに対する個別の予防策と低減策を指す場合もある。
    - 敵対的機械学習：*A Taxonomy and Terminology of Attacks and Mitigations* [3]（攻撃と低減の分類学と用語）」である。この報告書では、機械学習攻撃の主要な種類と攻撃者の目標、目的、能力、および攻撃の結果を低減し管理するための対応する方法を概説している。

NIST はまた、プロファイルの将来のバージョンにおいて、実装例の欄を追加することを検討している。**実施**例とは、あるタスクの一部または全部を達成する方法を提案する一文である。推奨事項と考慮事項」欄が "What "を記述するのに対し、「実施例」欄は "How "の選択肢を記述するものである。このプロファイルに追加される実施例は、SSDF version 1.1 で既に定義されている実施例を補足するものである。読者への注記

その他の参考資料や実施例に関する意見提供の詳細については、こちらを参照のこと。

注：このプロファイルは，SSDF version 1.1 [6]が既に含んでいるものを補足するものであり，単独で使用するのではなく，SSDF version 1.1 [6]と組み合わせて使用することを意図している。注意事項として，データガバナンスと管理のライフサイクルのほとんどの部分と同様に，AI モデルによる AI システムの展開と運用は，このプロファイルの適用範囲外である。

SSDF のプラクティスとタスクのナンバリングにはずれがある。例えば、PW.4 の練習には 3 つのタスクがある：PW.4.1、PW.4.2、PW.4.4 である。PW.4.3 は SSDF バージョン 1.0 のタスクであったが、バージョン 1.1 では別の場所に移動されたため、その ID は再利用されなかった。

表 1. AI モデル開発のための SSDF コミュニティ・プロフィール

プラクティス	タスク	優先順位	AI モデル開発に特有の推奨事項[R]、考慮事項[C]、注意事項[N]について	参考文献
組織（PO）を準備する				
ソフトウエア開発のセキュリティ要件を定義する（PO.1）：これにより、ソフトウエア開発ライフサイクル（SDLC）全体を通じてセキュリティ要 件を考慮することができ、また、要件情報を一度収集すれば共有できるため労力の重複を最小 限に抑えることができる。これには、内部的な情報源（例えば、組織の方針、ビジネス目標、リスクマネジメント戦略）と外部的な情報源（例えば、適用法や規制）からの要求事項が含まれる。	**PO.1.1：組織の**ソフトウエア開発インフラとプロセスに対す るすべてのセキュリティ要件を特定し文書化する。	高	**R1:** ソフトウエア開発インフラ及びプロセスのセキュリティ要件に AI モデル開発を含める **R2:**サイバーセキュリティ、プライバシー、再現性のための推奨プラクティスに従って、適切な AI モデルアーキテクチャとトレーニング技術を特定し、選択する。	**AI RMF：**マップ 1.3、1.5、1.6
	PO.1.2：組織が開発したソフトウエアのセキュリティ要件をすべて識別して文書化し、その要件を満たすようにする。	高	**R1：**組織の方針は、組織が開発したソフトウェアの AI モデル開発セキュリティに特化した現行の要件をすべてサポートするものとする。これらの要件には、AI モデル開発、AI モデル運用、データサイエンスの分野を含めるべきである。要件は、法律、規制、契約、標準など、多くの情報源からもたらされる。 **C1：**組織の既存のデータ格付方針およびプロセスを再利用または拡張することを検討する。	**AI RMF：**ガバナンス 1.1、1.2、3.2、4.1、5.1、6.1、マップ 1.1

プラクティス	タスク	優先順位	AI モデル開発に特有の推奨事項[R]、考慮事項[C]、注意事項[N]について	参考文献
			N1 : AI モデルの文書化には、データ、モデル、システムカードなどが考えられる。	
	PO.1.3:組織独自のソフトウェアで使用するための商用ソフトウェアコンポーネントを組織に提供するすべてのサードパーティに、要件をコミュニケーションする。[SSDF1.1 より修正]	中	**R1: サ**ードパーティ製ソフトウェアコンポーネントのコミュニケーション要件に、AI モデル開発のセキュリティを含める。	**AI RMF :** マップ 4.1、4.2 **OWASP:** LLM05-1
役割と責任を実施する（PO.2）: SDLC に関与する組織内外のすべての人が、SDLC 全体を通じて SDLC 関連の役割と責任を果たせるように準備されていることを確認する。	**PO.2.1 :** SDLC のすべての部分を網羅するために、必要に応じて新しい役割を作成し、既存の役割の責任を変更する。定義された役割と責任を定期的に見直し、維持し、必要に応じて更新する。	高	**R1:** SDLC 全体を通じて、SDLC 関連の役割と責任に AI モデル開発のセキュリティを含める。役割と責任には、AI モデルの開発、AI モデルの運用、データサイエンスが含まれるべきであるが、これらに限定されない。 **N1:**AI システム・プロデューサー、AI モデル・プロデューサー、その他のサードパーティ・プロバイダーに関わる役割と責任は、契約書に文書化することができる。	**AI RMF :** ガバナンス 2.1
	PO.2.2 : PO.2.2: セキュリティの発展に貢献する責務を担うすべての職員に、役割に応じたトレーニングを提供する。要員の習熟度と役割ベースのトレーニングを定期的に見直し、必要に応じてトレーニングを更新する。	高	**R1 :** 役割に応じたトレーニングには、AI モデルに対するサイバーセキュリティの脆弱性と脅威、およびそれらの可能な低減策を理解することが含まれるべきである。	**AI RMF :** ガバナンス 2.2 **OWASP: :**LLM04-7
	PO.2.3 : 上層部または認可担当官から、安全な開発に対するコミットメントを得る。	中	**R1:** リーダーシップは、AI モデルを含む安全な開発手法にコミットすべきである。	**AI RMF :** ガバナンス 2.3
支援ツールチェーンを実装する（PO.3）: 自動化を使用して、人間の労力を削減し、SDLC 全体を通じてセキュリティ対策の正確	**PO.3.1 :** 識別されたリスクを軽減するために、どのツールまたはツールタイプを各ツールチェーンに含め	高	**R1 :** 特に AI モデルで頻繁に使用される規模において、AI モデル開発を安全にし、人的労力を削減する自動化ツールチェーンの開発と実装を計画する。	**AI RMF :** メジャー 2.1 **OWASP:** LLM08

プラクティス	タスク	優先順位	AI モデル開発に特有の推奨事項[R]、考慮事項[C]、注意事項[N]について	参考文献
性、再現性、使いやすさ、包括性を改善する。ツールチェインとツールは、組織全体またはプロジェクト固有のように、組織のさまざまなレベルで使用することができ、また、構築パイプラインのように、SDLC の特定の部分に対処することができる。	なければならないか、または含めるべきかを明記する。		**N1 :** 理想的には、自動化されたツールチェーンが、AI モデルの安全な開発に関連する作業の大部分を行うことである。 **N2:** 工具の種類については、PO.4, PO.5, PS, PW を参照のこと。	
	PO.3.2 : ツールおよびツールチェーンを導入、運用、保守するために推奨されるセキュリティプラクティスに従う。	高	**R1:** AI モデル開発の安全性を確保し、特に AI モデルで頻繁に使用される規模において、人的労力を削減する自動化ツールチェーンを開発・導入する計画を実行する。 **R2:**リスクに見合った頻度でツールチェーンのセキュリティを検証する。	**AI RMF :** メジャー 2.1 **OWASP:** LLM05-3、LLM05-9、LLM08、LLM09
	PO.3.3 : PO.3.3: 組織で定義された安全なソフトウェア開発プラクティスをサポートする成果物を生成するようにツールを設定する。	中	**N1 :** アーティファクトとは「証拠の一部」である［16］。証拠とは、「信念や不信の根拠、証明の根拠となるデータ、真実や虚偽を立証するためのデータ」［17］である。成果物は、安全なソフトウェア開発の実践の記録を提供する。AI モデル開発に特有の成果物の例としては、トレーニングデータセットの完全性と出所の証明などがある。	**AI RMF :** メジャー 2.1
ソフトウェアのセキュリティチェックの基準を定義し、使用する（PO.4）: 開発中にソフトウェアのセキュリティをチェックするための基準を定義し、使用することで、SDLC か	**PO.4.1 :** ソフトウェアのセキュリティチェックの基準を定義し、SDLC 全体を通じて追跡する。	中	**R1:** 従来の SDLC を超えて、AI の開発ライフサイクル全体を通じてガードレールやその他の管理策を導入する。 **C1:** リスクベースの閾値を超えるソフトウエアセキュリティチェックについて、人間による レビューと承認を要求することを検討する。	**AI RMF :** 対策 2.3、2.7、管理 1.1 **OWASP:** LLM01-2

プラクティス	タスク	優先順位	AI モデル開発に特有の推奨事項[R]、考慮事項[C]、注意事項[N]について	参考文献
ら得られるソフトウェアが組織の期待に適合することを確実にする。	**PO.4.2 :** PO.4.2: 基準の裏付けとなる必要な情報を収集し、保護するためのプロセスや仕組みなどを導入する。	低	SSDF1.1 への追加はない	**AI RMF :** 対策 2.3、2.7、管理 1.1 **OWASP:** LLM01-2
セキュアな実装と維持 **ソフトウェア開発環境（PO.5）:** ソフトウェア開発環境のすべてのコンポーネントが内部および外部の脅威から強力に防御され、環境またはその中で開発・維持されているソフトウェアが危険にさらされないようにする。ソフトウェア開発環境の例としては、開発環境、AI モデル学習環境、ビルド環境、テスト環境、配布環境などがある。[SSDF1.1 より改変]。	**PO.5.1 :** ソフトウェア開発に関わる各環境を分離し、保護する。	高	**C1 :** 分離、セグメンテーション、封じ込め、API 経由のアクセス、その他の手段など、実行可能な範囲で実行環境を互いに分離することを検討する。 **R1:** モデル開発中の AI モデルユーザーのリソース使用量とレートを監視、追跡、制限する。 **R2:**本番データを含め、AI モデル開発中に使用される機密データは、組織が承認した環境およびその環境内の場所にのみ保存する。**R3:**最小特権の原則に従って、環境内のすべてのトレーニングパイプライン、モデル登録、およびその他のコンポーネントを防御する。 **R4:** パイプラインにおけるトレーニング関連のアクティビティと、モデルレジストリにおけるモデルの修正を継続的に監視する。 **R5:** 各環境を安全に構成するための推奨プラクティスに従うこと。 **R6:** 平文の秘密がないか、各環境を継続的に監視する。	**OWASP:** LLM01-1、LLM01-4、LLM04、LLM08、LLM10
	PO.5.2 : 開発エンドポイント（ソフトウェア設計者、開発者、テスト実施者、構築者など が開発タスクを実行するためのエンドポイント）を、リスクに応じたアプローチで保護し、堅牢化する。	中	SSDF1.1 への追加はない	**OWASP:** LLM01-1、LLM05-3、LLM05-9、LLM08

プラクティス	タスク	優先順位	AI モデル開発に特有の推奨事項[R]、考慮事項[C]、注意事項[N]について	参考文献
	PO.5.3 : ソフトウェア開発環境におけるソフトウェアの実行パフォーマンスと動作を継続的に監視し、潜在的な不審な活動やその他の問題を特定する。[SSDF1.1 には含まれていない。]	高	**R1 :** AI モデルや関連リソース（モデル API、重み、設定パラメータ、訓練データセットなど）をホストするすべての開発環境コンポーネントについて、セキュリティの継続的なモニタリングを実施する。 **R2:**継続的な監視・分析ツールは、AI モデルが関与する検知活動がリスク閾値を超えたとき、またはその他の追加調査に値するときにアラートを生成すべきである。	**AI RMF :** メジャー 2.4 **OWASP:** LLM03-7、LLM04、LLM058、LLM09、LLM10
防御ソフトウェア (PS)				
コードとデータのあらゆる形式を不正アクセスや改ざんから守る（PS.1）: コードとデータに対する不正な変更を、意図的であるか不注意であるかにかかわらず、防止する。一般に公開されることを意図していないコードやデータについては、ソフトウェアの盗難を防止するのに役立ち、また、攻撃者がそのようなコードやデータを見つけるのをより困難にしたり、時間をかけたりすることができる。 ソフトウェアの脆弱性を修正する。[SSDF1.1 より修正]。	**PS.1.1 :** ソースコード、実行可能コード、コンフィギュレーション・アズ・コードを含むすべての形式のコードを、最小権限の原則に基づいて保管し、認可された要員、ツール、サービスなどのみがアクセスできるようにする。	高	**R1:** 安全なコードストレージには、AI モデル、モデル重み、パイプライン、報酬モデル、および機密性、完全性、可用性の保護が必要なその他の AI モデル要素を含めるべきである。これらの要素は、すべて同じ場所に、または同じタイプのメカニズムを通じて保存される必要はない。 **R2:**最小特権の原則に従い、AI モデルやモデル要素の保存場所や実行場所にかかわらず、直接アクセスを最小限に抑える。 **R3:**報酬モデルを AI モデルやデータとは別に保管する。 **R4:** モデル重みへの間接的なアクセスのみを許可する。 **C1 :** モデルの重みに人間がアクセスできないようにすることを検討する。 **C2:**すべての AI モデル開発は、組織が承認した環境でのみ行うことを義務付けることを検討する。	**OWASP:** LLM10

プラクティス	タスク	優先順位	AI モデル開発に特有の推奨事項[R]、考慮事項[C]、注意事項[N]について	参考文献
	PS.1.2 : すべてのトレーニング、テスト、微調整、およびアライメントデータを、不正なアクセスや改変から防御する。[SSDF1.1 には含まれない。]	高	**R1 :** トレーニングデータ、テストデータ、微調整データ、アライメントデータの機密性（非公開データのみ）と完全性を継続的に監視する。 **C1:** 可能であれば、トレーニング、テスト、微調整、アライメントのデータを、将来の使用や参照のために安全に保管することを検討する。	**OWASP:** LLM03、LLM06、LLM10
	PS.1.3 : すべてのモデル重みと構成パラメータデータを、不正なアクセスや改変から保護する。[SSDF1.1 には含まれない。]	高	**R1:** モデルの重みと構成パラメータを、トレーニング、テスト、微調整、データ整列とは別に管理する。 **R2:**モデルの重みと構成パラメータの機密性（クローズドモデルのみ）と完全性を継続的に監視する。**R3:**最小特権の原則に従い、開発中の AI モデル重み、設定パラメータ、サービスへのアクセスを制限する。 **R4 :** 暗号化など、モデルウェイトに関連する、リスクに比例した追加のサイバーセキュリティ慣行を定め、実施する、 暗号ハッシュ、デジタル署名、マルチパーティ認可、エアギャップ環境などだ。	**OWASP: LLM10**
ソフトウェアリリースの完全性を検証する仕組みを提供する（PS.2）: ソフトウェアの取得者が、取得したソフトウェアが正当なものであり、改ざんされていないことを確認できるようにする。	**PS.2.1 :** ソフトウェアの完全性検証情報をソフトウェア取得者が利用できるようにする。	中	**R1:** AI モデルとそのコンポーネント、成果物、文書の暗号ハッシュまたはデジタル署名を生成し、提供する。 **R2:**AI モデルの変更にデジタル署名を提供する。	**OWASP:** LLM05-6

プラクティス	タスク	優先順位	AI モデル開発に特有の推奨事項[R]、考慮事項[C]、注意事項[N]について	参考文献
各ソフトウェアリリースをアーカイブし、防御する（PS.3）：リリース後のソフトウェアに発見された脆弱性を識別、分析、排除するために、ソフトウェアリリースを保存する。	**PS.3.1：**各ソフトウェアリリースに必要なファイルおよびサポートデータ（完全性検証情報、出所データなど）をセキュアにアーカイブする。	低	**R1:** データセット作成とモデル学習をサポートするインフラツール（事前処理、変換、収集など）のバージョン管理とトラッキングを行う。 **R2:**AI モデル選択の正当性を示す文書を保持情報に含める。 **R3:** データの前処理やモデル構築など、トレーニングプロセス全体の文書を含める。 **N1：**AI モデルとそのコンポーネントは、現時点で組織の資産目録に追加する必要があるかもしれない。	**OWASP:** LLM10
	PS.3.2：各ソフトウェアリリースのすべてのコンポーネント（ソフトウェア部品表[SBOM]など）の出所データを収集し、保護し、維持し、共有する。 ソフトウェア成果物のサプライチェーン・レベル [SLSA])である。[SSDF 1.1]から変更された。	中	**R1：**モデルの構築に使用されたトレーニングライブラリ、フレームワーク、パイプラインを含め、AI モデルとそのコンポーネント、派生物の出所を追跡する。 **R2:**センシティブなデータ（例：ペイメントカードデータ、保護された医療情報、個人を特定できるその他の種類の情報）でトレーニングされた AI モデルを追跡し、モデルへのアクセスを、トレーニングに使用されたセンシティブなデータに既にアクセスできる個人に制限すべきかどうかを判断する。**C1：**AI モデルに使用されるトレーニング、テスト、微調整、調整データの出所を開示することを検討する。	**OWASP:** LLM03-1、LLM05-4、LLM05-5、LLM10
安全性の高いソフトウェアを作る (PW)				
セキュリティを満たすソフトウェアを設計する	**PW.1.1：**脅威モデリング、攻撃モデリング、攻撃対象領域マッピングなどのリスクモデリングを使用し	高	**R1：**関連する AI モデル特有の脆弱性と脅威の種類をリスクモデリングに組み込む。これらの脆弱性と脅威の例としては、学習データのポイズニング、入力と出力に含	**AI RMF：**ガバナンス 4.1、4.2、マップ

プラクティス	タスク	優先順位	AI モデル開発に特有の推奨事項[R]、考慮事項[C]、注意事項[N]について	参考文献
要件とセキュリティリスクを低減する **（PW.1）：**運用中にソフトウエアが直面しそうなセキュリティリスクと、ソフトウエアの設計とアーキテ クチャがそれらのリスクをどのように低減すべきかを決定する。リスクに基づく分析によって、セキュリ ティ要件を緩和または免除すべきことが示された場合は、それを正当化する。ソフトウエア設計時にセキュリティ要件とリスクに対処すること（セキュア・バイ・デザ イン）は、ソフトウエアセキュリティ改善の鍵であり、開発効率の向上にも役立つ。	て、ソフトウェアのセキュリティリスクをアセスメントする。		まれる悪意あるコードやその他の不要なコンテンツ、敵対的なプロンプトから生じるサービス拒否状態、サプライチェーン攻撃、不正な情報開示、AI モデルの重みの窃取、データパイプラインの誤設定などがある。[3] **C1：**将来の AI モデルのバージョンおよび AI モデルリリース後のデリバティブについて、定期的なリスクモデリングの更新を検討する。 **C2:**リスクモデリング中に、AI モデルがクリティカルパスに入っておらず、人間がループに入らずに重大なセキュリ ティ上の意思決定を行っていないことを確認する。	5.1、対策 1.1、管理 1.2、1.3 **OWASP:** LLM01、LLM02、LLM03、LLM04、LLM05、LLM06、LLM07、LLM08、LLM09、LLM10
	PW.1.2：PW.1.2: ソフトウエアのセキュリティ要件、リスク、設計上の決定を追跡し、維持管理する。	中	SSDF1.1 への追加はない	**AI RMF：**ガバナンス 4.1、4.2、マップ 2.1、2.2、2.3、3.2、3.3、4.1、4.2、5.2、管理 1.2、1.3、1.4
	PW.1.3：セキュリティ機能やサービスを独自に実装するのではなく、標準化されたセキュリティ機能やサービス（例えば、既存のログ管理、ID 管理、アクセス制御、脆弱性管理システムとの統合を可能にする）を使用するためのサポートを、適切な場合に組み込む。	中	SSDF1.1 への追加はない	
ソフトウェア設計を検証する **セキュリティ要件とリスク情報への適合** **（PW.2）：**ソフトウェアがセキュリティ要件	**PW.2.1：**1)設計に関与していない有資格者（複数可）、及び 2)ツールチェーンにインス タンス化された自動プロセスにソフトウエア設計をレビューさせ、	高	SSDF1.1 への追加はない	**AI RMF：**対策 2.7、管理 1.1

プラクティス	タスク	優先順位	AI モデル開発に特有の推奨事項[R]、考慮事項[C]、注意事項[N]について	参考文献
を満たし、識別されたリスク情報に十分に対処することを保証する。	ソフトウ エア設計がすべてのセキュリティ要件を満たし、個人を特定できるリ スク情報に十分に対処していることを確認し、実施させる。[SSDF1.1]を修正した。			
トレーニング、テストの完全性を確認する、 **使用前のデータの微調整と整合（PW.3）:** AI モデルのトレーニング、テスト、微調整、整列の一環として、AI モデルのサイバーセキュリティに悪影響を及ぼす可能性のあるデータが消費されないようにする。[SSDF1.1 には含まれていない。]	**PW.3.1 :** PW3.1:AI モデルのトレーニング、テスト、微調整、整列の目的でデータを使用する前に、データ・ポイズニング、バイアス、均質性、改ざんの兆候についてデータを分析し、必要に応じてリスクを軽減する。[SSDF1.1 には含まれない。]	高	**R1 :** 使用前に、トレーニング、テスト、微調整、アライメントデータの出所（既知の場合）と完全性を検証する。 **R2:**AI モデルのトレーニング、テスト、ファインチューニング、アライメントデータを分析し、変更するための適切な方法を選択し、適用する。手法の例としては、異常検知、バイアス検知、データクリーニング、データキュレーション、データフィルタリング、データサニタイゼーション、ファクトチェック、ノイズ除去などがある。 **C1:** 探索的データ分析技術[18]のように、データを調査するために人間を使うことを検討する。	**AI RMF :** 対策 2.1、管理 1.2、1.3 **OWASP:** LLM03、LLM06
	PW.3.2 : PW3.2 : AI モデルに使用されるすべてのトレーニングデータ、テストデータ、微調整データ、アライメントデータの出所がわかっている場合はそれを追跡し、どのデータの出所がわからないかを文書化する。[SSDF1.1 には含まれていない。]	中	**N1 :** 実績は常に知られているわけではないので、すべてのケースで検証できるわけではない。しかし、可能な限り実績の追跡と検証を行い、実績が不明な場合は追跡を行うことは、セキュリティ上有益である。	**AI RMF :** メジャー 2.1 **OWASP:** LLM03-1 **アドバンス ML**
	PW.3.3 : 攻撃防御を改善するために、訓練データとテストデータに敵対的なサンプルを含める。[SSDF1.1 には含まれていない。]	中	**R1:** 敵対的サンプルをテストするプロセスと対応するコントロールを使用し、トレーニングとテストの使用に適切なガードレールを設ける。	**OWASP:** LLM03-6、LLM05-7 **アドバンス** ML

プラクティス	タスク	優先順位	AI モデル開発に特有の推奨事項[R]、考慮事項[C]、注意事項[N]について	参考文献
可能であれば既存の安全性の高いソフトウェアを再利用する（PW.4）：ソフトウエアの開発コストを低減し、ソフトウエア開発を迅速化し、セキュリ ティ体制をチェック済みのソフトウエアモジュールやサービスを再利用する ことで、ソフトウエアに新たなセキュリティ脆弱性が混入する可能性を低減する。これは、暗号モジュールやプロトコルのようなセキュ リティ機能を実装するソフトウエアにとって特に重要である。	**PW.4.1：PW4.1：**組織のソフトウェアで使用するために、商用、オープンソース、その他のサードパーティ開発者から、安全性の高いソフトウェアコンポーネント（ソフトウェアライブラリ、モジュール、ミドルウェア、フレームワークなど）を取得し、維持する。	中	**C1：**新しい AI モデルを作るのではなく、既存の AI モデルを使うことを検討する。	**OWASP:** LLM05
	PW.4.2：サードパーティ製のソフトウェアコンポーネントでは対応できない、社内の一般的なソフトウェア開発のニーズを満たすために、SDLC プロセスに従って、安全性の高いソフトウェアコンポーネントを社内で作成し、維持する。	低	SSDF1.1 への追加はない	
	PW.4.4：取得した商用ソフトウェア、オープンソースソフトウェア、その他すべてのサードパーティソフトウェアコンポーネントが、そのライフサイクルを通じて、組織が定義した要件に準拠していることを検証する。	高	**R1:** 既存の AI モデル、またはその他の取得した AI コンポーネント（トレーニング、テスト、微調整、データセットの調整、報酬モデル、適応レイヤー、設定パラメータを含む）の完全性、出所、セキュリティを、使用前に検証する。 **R2:**取得した AI モデルやそのコンポーネントを使用する前に、脆弱性や悪意のあるコンテンツがないかスキャンし、徹底的にテストする。	**OWASP:** LLM05-2、LLM05-6 **アドバンス** ML
Secure Coding Practices（PW.5）に従ってソースコードを作成する：ソフトウエアに含まれるセキュリティ脆弱 性の数を減らし、ソースコード作成時に生じ る脆弱性のうち、組織で定義した脆弱 性の重大度基準を満たすか超えるものを 最小限に抑えることでコストを削減する	**PW.5.1：**PW5.1：組織の要件を満たすために、開発言語と環境に適したすべてのセキュアコーディングプラクティスに従う。	高	**R1:** 安全なコーディングの実践を拡大し、AI 技術固有の考慮事項を含める。 **R2:**入力（プロンプトやユーザーデータを含む）と出力の取り扱いを慎重にコーディングする。すべての入力と出力は、AI モデルのコンテキスト内でログに記録され、分析され、検証されるべきであり、問題のあるものはサニタイズされるか、削除されるべきである。	**AI RMF：**1.2、1.3、1.4 を管理する **OWASP:** LLM01、LLM02、LLM04-1、LLM06、LLM07、LLM09-9、LLM10

プラクティス	タスク	優先順位	AI モデル開発に特有の推奨事項[R]、考慮事項[C]、注意事項[N]について	参考文献
			R3:不正なコードの実行を防ぐため、入出力を暗号化する。	
コンパイル、インタープリタ、ビルドプロセスを設定して、実行可能ファイルのセキュリティを改善する。(PW.6)：テスト実施前に脆弱性を排除することで、ソフトウエアのセキュリティ脆弱性の数を減らし、コストを削減する。	**PW.6.1：**コンパイラ、インタプリタ、ビルドツールのうち、実行ファイルのセキュリティを改善する機能を持つものを使用する。	低	**C1：**悪意のあるコンテンツの序文を減らす、または排除する、安全なモデル・シリアライゼーション・メカニズムの使用を検討する。	
	PW.6.2：どのコンパイラ、インタプリタ、ビルドツールの機能を使用し、それぞれをどのように設定すべきかを決定し、承認された設定を実装して使用する。	低	**C1:** コンパイラ、インタープリタ、ビルドツールのバージョンと機能を、実績追跡の一部として取り込むことを検討する。	
脆弱性を識別し、セキュリティ要件への準拠を検証するために、人間が読 み取り可能なコードをレビュー及び／又は分析する間が読むことができるものを見直し、分析する **(PW.7)：**ソフトウェアがリリースされる前に脆弱性を修正し、悪用されないようにする。自動化された手法を使用することで、脆弱性の検知に必要な労力とリソースを削減できる。人間が読み取り可能なコードには、ソースコード、スクリプト、その他組織が人間が読み取り可能と判断するあらゆる形式のコードが含まれる。	**PW.7.1：PW7.1：**コードレビュー（人が直接コードを見て問題を発見する）とコード分析（完全自動化された方法で、または人と協力して、コード内の問題を発見するためにツールを使用する）のどちらを使用するかを、組織の定義に従って決定する。	中	**R1：**コードレビューと分析の方針またはガイドラインは、AI モデルやその他の関連コンポーネントのコードを含むべきである。 **C1：**AI モデルのテストに加えて、AI モデルのコードのスキャンを行うことを検討する。	
	PW.7.2：組織のセキュアコーディング標準に基づいてコードレビューおよび/またはコード分析を実施し、発見されたすべての問題と推奨される改善策を、開発チームのワークフローまたは課題追跡システムに記録し、トリアージする。	高	**R1：**組織のコードレビューおよび分析ポリシーまたはガイドラインに従って、すべての AI モデルにマルウェア、脆弱性、バックドア、その他のセキュリティ上の問題がないかどうかを分析する。	**AI RMF：**対策 2.3、2.7、管理 1.1、1.2、1.3、1.4 **OWASP:** LLM03-7d、LLM07-4
実行可能コードをテストして識別する	**PW.8.1：**PW 8.1：以前のレビュー、分析、またはテストによって識別されなかった脆弱性を発見するために、実行可能コードテストを実施すべきかどうか、	高	**R1：**コードテストの方針とガイドラインに AI モデルを含める。ユニットテスト、統合テスト、侵入テスト、レッドチーム、ユースケーステスト、敵対的テストなど、	

プラクティス	タスク	優先順位	AI モデル開発に特有の推奨事項[R]、考慮事項[C]、注意事項[N]について	参考文献
脆弱性を特定し、セキュリティ要件への準拠を検証する（PW.8）: 脆弱性の特定を支援し、悪用されないようにソフトウエアのリリース前に脆弱性を修正できるようにする。自動化された手法を使用することで、脆弱性の検知に必要な労力とリソースを削減し、トレーサビリティと再現性を向上させる。実行可能なコードには、バイナリ、直接実行されるバイトコード、ソースコード、および組織が実行可能とみなすその他の形式のコードが含まれる。	また実施する場合はどのような種類のテストを使用すべきかを決定する。		いくつかの形式のコードテストを AI モデルに使用することができる。 **C1:** 可能であれば、回帰テストの一環として、開発パイプライン内でのテストの自動化を検討する。	
	PW.8.2 : テストのスコープを設定し、テストを設計し、テストを実施し、その結果を文書化する。	高	**R1 :** 組織のコードテスト方針またはガイドラインに従って、すべての AI モデルの脆弱性をテストする。 **R2:**AI モデルが再トレーニングされたり、新しいデータソースが追加されたときに再テストする。	**AI RMF :** 対策 2.2、2.3、2.7 ; 管理 1.1、1.2、1.3、1.4 **OWASP: である :** LLM03-7d、LLM05-7、LLM07-4
ソフトウェアをデフォルトで安全な設定にする（PW.9）: インストール時にソフトウエアのセキュリティを改善し、ソフトウエアが脆弱なセキュリティ設定のまま配備され、危険にさらされる可能性を低減する。	**PW.9.1 :** セキュリティに影響を及ぼす各設定、またはセキュリティに関連する設定を、デフォルト設定が安全で、プラットフォー ム、ネットワークインフラ、またはサービスが提供するセキュリティ機能を弱めないように構成する方法を決定する ことで、安全なベースラインを定義する。	中	SSDF1.1 への追加はない	**AI RMF :** メジャー 2.7
	PW.9.2 : デフォルト設定を実行する (該当する場合は、デフォルト設定のグループ)、およびソフトウェア管理者のために各設定を文書化する。	中	**N1 :** 設定の文書化は、セキュアなベースラインを定義するとき（PW.9.1 参照）など、プロセ スの早い段階で行うことができる。	**AI RMF :** 対策 2.7、管理 1.2、1.3、1.4
脆弱性への対応（RV）				

プラクティス	タスク	優先順位	AI モデル開発に特有の推奨事項[R]、考慮事項[C]、注意事項[N]について	参考文献
脆弱性の識別と確認を継続的に行う（RV.1）：脆弱性をより迅速に特定し、リスクに応じてより迅速に是正できるようにし、攻撃者の隙を減らす。	**RV.1.1：**ソフトウェア及びソフトウェアが使用するサードパーティ製コンポーネントの潜在的な脆弱性について、ソフトウェア取得者、ユーザ及び一般情報源から情報を収集し、信頼できるすべての報告を調査する。	高	**R1:** AI モデルのすべての入出力をログに記録し、監視し、分析し、セキュリティ上及び性能上の問題の可能性を検知する（PO.5.3 を参照）。 **R2:**AI モデルの利用者に、潜在的なセキュリティやパフォーマンスの問題を報告する仕組みを認識させる。 **N1：ここでいう**「ユーザー」とは、AI モデルを利用する AI システム製作者や買収者を指す。 **R3:**脆弱性データベースやインシデントデータベースを監視し、AI モデルの構築に使用される機械学習フレームワークやライブラリなど、AI 関連の懸念事項に関する情報を入手する。	**AI RMF:** ガバナンス 4.3、5.1、6.1、6.2; 対策 1.2、2.4、2.5、2.7、3.1、3.2、3.3; 管理 4.1 **OWASP: :** LLM03-7a, LLM09, LLM10
	RV.1.2：RV.1.2：ソフトウェアのコードをレビュー、分析、テストし、これまで検出されていなかった脆弱性の存在を特定または確認する。	中	**R1：**AI モデルを頻繁にスキャンしてテストし、これまで検出されていなかった脆弱性を特定する。 **R2:**継続的なスキャンとテストは主に自動化に依存し、必要に応じて人間が関与する。 **R3:**AI モデルの定期的な監査を実施する。	**AI RMF：**ガバナンス 4.3、対策 1.3、2.4、2.7、3.1、管理 4.1 **OWASP:** LLM03-7b, LLM03-7d
	RV.1.3：脆弱性の開示と是正に取り組む方針を持ち、その方針をサポートするために必要な役割、責任、プロセスを実施する。	中	**R1：**AI モデルの脆弱性を、組織の脆弱性情報開示および是正方針に含める。 **R2:**AI モデルの利用者に、AI モデル固有の限界と、サイバーセキュリティ上の問題に遭遇した場合の報告方法を認識させる。	**AI RMF：**ガバナンス 4.3、5.1、6.1、測定 3.1、3.3、管理 4.3

プラクティス	タスク	優先順位	AI モデル開発に特有の推奨事項[R]、考慮事項[C]、注意事項[N]について	参考文献
脆弱性をアセスメントし、優先順位を付け、是正する（RV.2）：脆弱性をリスクに応じて確実に是正し、攻撃者の攻撃機会を減らす。	**RV.2.1：**RV.2.1：各脆弱性を分析し、その是正またはその他のリスク対応を計画するために必要なリスクに関する十分な情報を収集する。	中	**N1：**これには、生成的 AI やデュアルユースファウンデーションモデルの入出力を深く分析し、正常な動作からの逸脱を検知することも含まれる。	**AI RMF：**ガバナンス 4.3、5.1、6.1、対策 2.7、3.1、管理 1.2、2.3、4.1 **アドバンス** ML
	RV.2.2：脆弱性に対するリスク対応を計画し、実施する。	高	**R1：**AI モデルのリスク対応は、再構築に伴う時間と費用を考慮すべきである。 **R2:**AI モデルの使用をいつ中止し、いつ以前のバージョンとそのコンポーネントにロールバックするかの基準とプロセスを確立し、実施する。 **C1：**AI モデルのリスクが十分に解決されるまでは、いつでも AI モデルの使用を中止し、別の手段で業務を継続できるように準備しておくことを検討する。	**AI RMF：**ガバナンス 5.1、5.2、6.1、対策 3.3、管理 1.3、2.1、2.3、2.4、4.1
脆弱性を分析し、その根本原因を特定する（RV.3）：将来、脆弱性が発生する頻度を減らすことに貢献する。	**RV.3.1：**識別された脆弱性を分析し、その根本原因を特定する。	中	**N1：**トレーニング、テスト、微調整、データ調整などを事後的に見直すことで、根本的な原因を特定することができる。	**AI RMF：**ガバナンス 5.1、6.1、対策 2.7、3.1、管理 2.3、4.1
	RV.3.2：RV.3.2：根本原因を長期的に分析し、特定のセキュアコーディング手法が一貫して守られていないなどのパターンを特定する。	中	SSDF1.1 への追加はない	**AI RMF：**ガバナンス 5.1、6.1、対策 2.7、3.1、管理 4.1、4.3
	RV.3.3：外部からの報告を待つのではなく、類似の脆弱性がないかソフトウェアをレビューし、脆弱性のクラスを根絶し、積極的に修正する。	中	SSDF1.1 への追加はない	**AI RMF：**ガバナンス 5.1、5.2、6.1、対策 2.7、3.1、管理 4.1、4.2、4.3

プラクティス	タスク	優先順位	AI モデル開発に特有の推奨事項[R]、考慮事項[C]、注意事項[N]について	参考文献
	RV.3.4 : SDLC プロセスをレビューし、ソフトウェアのアップデートや新しいソフトウェアの作成時に根本原因が再発しないように（または再発する可能性を低減するように）、適切であれば更新する。	中	SSDF1.1 への追加はない	**AI RMF :** ガバナンス 5.2、6.1、対策 2.7、3.1、管理 4.2、4.3

参考文献

[1] 大統領令 14110（2023 年） 安全、安心、信頼できる人工知能の開発と利用。(ホワイトハウス、ワシントン DC)、DCPD-202300949、2022 年 10 月 30 日。https://www.govinfo.gov/app/details/DCPD-202300949。

[2] 国立標準技術研究所（2023）人工知能リスクマネジメントフレームワーク（AI RMF 1.0)。(National Institute of Standards and Technology, Gaithersburg, MD） NIST Artificial Intelligence (AI) Report, NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1.

[3] Vassilev A, Oprea A, Fordyce A, Anderson H (2024) Adversarial Machine Learning : A Taxonomy and Terminology of Attacks and Mitigations.(National Institute of Standards and Technology, Gaithersburg, MD) NIST Artificial Intelligence (AI) Report, NIST AI 100-2e2023. https://doi.org/10.6028/NIST.AI.100-2e2023.

[4] Schwartz R, Vassilev A, Greene K, Perine L, Burt A, Hall P (2022) Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.(National Institute of Standards and Technology, Gaithersburg, MD), NIST 特別刊行物 (SP) 1270. https://doi.org/10.6028/NIST.SP.1270.

[5] NIST (2024) Dioptra. (National Institute of Standards and Technology, Gaithersburg, MD.) Available at https://pages.nist.gov/dioptra/.

[6] Souppaya MP, Scarfone KA, Dodson DF (2022) Secure Software Development Framework (SSDF) Version 1.1 : ソフトウェアの脆弱性リスクを低減するための推奨事項。(National Institute of Standards and Technology, Gaithersburg, MD), NIST 特別刊行物 (SP) 800-218. https://doi.org/10.6028/NIST.SP.800-218.

[7] 大統領令 14028 号（2021 年）国家のサイバーセキュリティを改善する。(ホワイトハウス、ワシントン DC)、DCPD-202100401、2021 年 5 月 12 日。https://www.govinfo.gov/app/details/DCPD-202100401。

[8] National Institute of Standards and Technology (2024) The NIST Cybersecurity Framework (CSF) 2.0 (National Institute of Standards and Technology, Gaithersburg, MD) https://doi.org/10.6028/NIST.CSWP.29。

[9] 合同作業部会（2020 年） 情報システムおよび組織のためのセキュリティおよびプライバシー管理。(National Institute of Standards and Technology, Gaithersburg, MD），NIST 特別刊行物 (SP) 800-53, Rev. 5.2020 年 12 月 10 日時点の更新を含む。 https://doi.org/10.6028/NIST.SP.800-53r5

[10] Boyens JM, Smith AM, Bartol N, Winkler K, Holbrook A, Fallon M (2022) Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations.(National Institute of Standards and Technology, Gaithersburg, MD), NIST 特別刊行物 (SP) 800-161r1. https://doi.org/10.6028/NIST.SP.800-161r1.

[11] NIST (2023) NIST AI RMF Playbook.(国立標準技術研究所、メリーランド州ゲイサーズバーグ） https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook で入手できる。

[12] 国立標準技術研究所（2024）人工知能リスクマネジメントフレームワーク：生成的人工知能プロファイル。(National Institute of Standards and Technology, Gaithersburg, MD），NIST Artificial Intelligence (AI) Report, NIST AI 600-1.https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf。

[13] 国立標準技術研究所（2020）NIST プライバシーフレームワーク：エンタープライズリスクマネジメントを通じてプライバシーを改善するためのツール、バージョン 1.0。(National Institute of Standards and Technology, Gaithersburg, MD), NIST Cybersecurity White Paper (CSWP) NIST CSWP 10. https://doi.org/10.6028/NIST.CSWP.10.

[14] Stine KM, Quinn SD, Witte GA, Gardner RK (2020) Integrating Cybersecurity and Enterprise Risk Management (ERM).(National Institute of Standards and Technology, Gaithersburg, MD）、NIST Interagency or Internal Report (IR) 8286。 https://doi.org/10.6028/NIST.IR.8286。

[15] OWASP:（2023) OWASP: Top 10 for LLM Applications Version 1.1.https://llmtop10.com。

[16] Waltermire DA, Scarfone KA, Casipe M (2011) Specification for the Open Checklist Interactive Language (OCIL) Version 2.0.(National Institute of Standards and Technology, Gaithersburg, MD), NIST Interagency or Internal Report (IR) 7692。https://doi.org/10.6028/NIST.IR.7692。

[17] Ross RS, McEvilley M, Winstead M (2022) Engineering Trustworthy Secure Systems.(National Institute of Standards and Technology, Gaithersburg, MD), NIST 特別刊行物 (SP) NIST SP 800-160v1r1. https://doi.org/10.6028/NIST.SP.800-160v1r1.

[18] NIST/SEMATECH (2012) What is EDA?*Engineering Statistics Handbook*, eds Croarkin C, Tobias P (National Institute of Standards and Technology, Gaithersburg, MD), Section 1.1.1.https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm。

[19] Reznik L (2022) インテリジェント・セキュリティ・システム : How Artificial Intelligence, Machine Learning and Data Science Work For and Against Computer Security.(Wiley-IEEE Press.) https://ieeexplore.ieee.org/book/9562694 から入手できる。

附属書 A. 用語集

人工知能

人間が定義した特定の目的に対して、現実または仮想環境に影響を与える予測、推奨、または決定を行うことができる機械ベースのシステム。[1]

人工知能モデル

AI 技術を実装し、与えられた入力セットから出力を生成するために計算、統計、または機械学習技術を使用する情報システムのコンポーネント。[1]

人工知能レッドチーム

AI システムの欠陥や脆弱性を見つけるための構造化されたテスト作業で、多くの場合、管理された環境で、AI の開発者と協力して行われる。[1]

人工知能システム

データシステム、ソフトウェア、ハードウェア、アプリケーション、ツール、またはユーティリティの全部または一部が AI を使用して動作する。[1]

データサイエンス

データから意味のある洞察を抽出するために、領域の専門知識、プログラミングスキル、数学や統計の知識を組み合わせる分野である。[19]

財団兼用モデル

広範なデータで学習され、一般的に自己監視を使用し、少なくとも数百億のパラメータを含み、広範な文脈に適用可能であり、安全保障、国家経済安全保障、国家公衆衛生もしくは安全、またはそれらの組み合わせに重大なリスクをもたらすタスクにおいて、以下のような高いレベルのパフォーマンスを示すか、または示すように容易に修正できる AI モデル：

(i) 化学、生物、放射性物質、核兵器（CBRN）の設計、合成、入手、使用について、専門家以外の参入障壁を大幅に引き下げる；

(ii) 自動化された脆弱性発見と悪用により、サイバー攻撃の潜在的な標的を幅広く攻撃し、強力な攻撃的サイバー作戦を可能にする。

(iii) 欺瞞や難読化によって、人間の制御や監視から逃れることを可能にする。

たとえモデルが、ユーザが関連する安全でない機能を利用することを防ごうとする技術的セーフガードとともにエンドユーザにプロバイダされたとしても、この定義を満たす。[1]

生成的人工知能

派生合成コンテンツを生成するために、入力データの構造や特性をエミュレートする AI モデルのクラス。これには、画像、動画、音声、テキスト、その他のデジタル・コンテンツが含まれる。[1]

モデル重量

AI モデル内の数値パラメータで、入力に対するモデルの出力を決定するのに役立つ。[1]

証明

特定のデータの出所またはソースに関するメタデータ。[13]