仮訳

人工知能(AI)への取組

Australian Government
Australian Signals Directorate

ASD AUSTRALIAN SIGNALS DIRECTORATE
ACSC Australian Cyber Security Centre

National Cyber Security Centre
a part of GCHQ

Communications Security Establishment
Canadian Centre for Cyber Security

Centre de la sécurité des télécommunications
Centre canadien pour la cybersécurité

National Cyber Security Centre
PART OF THE GCSB

certnz

Federal Office for Information Security

INCD
Israel National Cyber Directorate

NISC
内閣サイバーセキュリティセンター
National center of Incident readiness and Strategy for Cybersecurity

NSM
NORWEGIAN NATIONAL CYBER SECURITY CENTRE

CSA SINGAPORE

NATIONAL CYBER SECURITY CENTRE SWEDEN

人工知能（AI）への取組

はじめに

本文書の目的は、AI システムをセキュアに利用する方法に関するガイダンスを、利用する組織（以下、「組織」）に提供することである。本文書は、AI システムに関連するいくつかの重要な脅威を要約し、組織がリスクを管理しながら AI に取り組むために可能な措置を検討するよう促す。また、自己ホスト型 AI システムとサードパーティーホスト型 AI システムを使用する組織の双方を支援するための緩和策を提供する。

本文書は、豪州通信電子局（ASD）豪州サイバーセキュリティセンター（ACSC）が、以下の国際的なパートナーと協力して作成した。

- 米国サイバーセキュリティ・インフラストラクチャー安全保障庁（CISA）、米国連邦捜査局（FBI）及び国家安全保障局（NSA）
- 英国国家サイバーセキュリティセンター (NCSC-UK)
- カナダサイバーセキュリティセンター (CCCS)
- ニュージーランド国家サイバーセキュリティセンター (NCSC-NZ)
- ドイツ連邦情報セキュリティ庁 (BSI)
- イスラエル国家サイバー総局 (INCD)
- 日本内閣サイバーセキュリティセンター (NISC）及び日本内閣府科学技術・イノベーション推進事務局
- ノルウェー国家サイバーセキュリティセンター (NCSC-NO)
- シンガポールサイバーセキュリティ庁 (CSA)
- スウェーデン国家サイバーセキュリティセンター

本文書のガイダンスは、セキュアな AI システムを開発することよりも、AI システムをセキュアに使用することに重点を置いている。署名組織は、AI システムの開発者に対し、セキュア AI システム開発ガイドライン（Guidelines for Secure AI System Development）を参照することを推奨する。

AI とは何か?

AI とは、視覚認識、言語認識、意思決定、言語間翻訳など、通常は人間の知性を必要とするタスクを実行できるコンピュータシステムの理論と開発である。現代の AI は通常、機械学習アルゴリズムを用いて構築される。

AI にはいくつかの重要な分野があり、以下を含むが、それらに限定されるものではない。

- **機械学習**は、人間によりプログラムされたルールを用いず、コンピュータがデータのパターンを認識し文脈（コンテキスト）をもたらすソフトウェアコンポーネント（モデル）を表す。機械学習アプリケーションは、統計的推論に基づいて予測、推奨、決定をすることができる。

- **自然言語処理**は、テキスト、画像、ビデオ、音声データ等の人間言語ソースから情報を分析し、導き出す。自然言語処理アプリケーションは、一般的に言語の分類と解釈に使用される。多くの自然言語処理アプリケーションは、自然言語を処理するだけでなく、自然言語を模倣したコンテンツを生成する。
- **生成 AI** は、テキスト、画像、音声、コード、その他のデータ様式などのコンテンツの新しい用例を生成するためにデータモデルを使用するシステムを指す。生成 AI アプリケーションは通常、大量の実世界におけるデータで学習され、限定的または非特定的なプロンプトであっても、それらのプロンプトから人間が生成したコンテンツを近似することができる。

AI システムは、世界的に最も急速に成長しているアプリケーションのひとつである。AI は、インターネット検索、衛星ナビゲーション、推奨システムを動かしている。また、AI は大規模なデータセットの整理、定型的な業務の自動化、創造的な取組、顧客対応、物流、医療診断、サイバーセキュリティをはじめとするビジネス活動の強化など、従前人間が行っていた活動を処理するために使用されることも増えている。

あらゆる分野の組織が、AI を活用して業務を改善する機会を模索している。AI は効率を高め、コストを削減する可能性を秘めている一方で、意図的または不注意に害をもたらす可能性もある。このため、政府、学術界及び産業界は、研究、規制、政策、ガバナンスなどを通じて、この技術に関連するリスクを管理する役割を担っている。

AI に取り組むにあたっての課題

全てのデジタルシステムと同様、AI は機会と脅威の両方をもたらす。AI の利点をセキュアに活用するためには、これらのシステムに関わる全ての利害関係者（プログラマー、ユーザー、管理職、アナリスト、マーケティング担当者など）が、どのような脅威が自分達に当てはまり、どのように軽減できるのかを理解するために時間を取る必要がある。

AI に関する脅威を以下のとおり説明する。これらの脅威の説明は、AI の利用を思いとどまらせるためではなく、全ての利害関係者が AI をセキュアに利用できるよう支援するためである。AI の脅威に対する安全確保のためのサイバーセキュリティ緩和策は、本文書の後半に掲載する。AI 特有の脅威、敵対者の戦術、リスク管理方法についての追加情報については、MITRE ATLAS と米国国立標準技術研究所（NIST）の AI リスク管理フレームワークを参照願いたい。

1. AI モデルのデータポイズニング

NIST は、機械学習のデータ駆動アプローチが、他システムとは異なるセキュリティとプライバシー上の課題をもたらすと説明する（NIST の「Adversarial Machine Learning：A Taxonomy and Terminology of Attacks and Mitigations」を参照願いたい）。これらの課題には、潜在的に、学習データの改ざんやモデル脆弱性の悪用（敵対的 AI としても知られる）が含まれ、機械学習のツールやシステムのパフォーマンスに悪影響を及ぼし得る。

敵対的改ざんの方法の一つにデータポイズニングがある。データポイズニングには、AI モデルの学習データを操作することが含まれる。そうすることにより、モデルが誤ったパターンを学習しデータを誤って分類したり、または、不正確、偏った、悪意のある出力を生成する可能性がある。AI システムの出力の完全性に依存する組織機能は、データポイズニングによって悪影響を受け得る。AI モデルの学習データは、新しいデータを挿入し、又は既存のデータを修正することで操作され得る。又は、学習データは、もともと侵害されていた情報源から取得されることもあり得る。データポイズニングは、モデルのファインチューニング・プロセスでも発生し得る。正常に機能しているか判断するためフィードバックを受け取る AI モデルは、質の低いフィードバックや誤ったフィードバックによって操作される可能性がある。

ケーススタディ：Tay チャットボットポイズニング

2016 年、マイクロソフトは機械学習を活用した「Tay」と呼ばれるツイッターのチャットボットを試用した。このチャットボットは、ユーザーとの会話を利用して自己学習し、機能を適応させるものであったため、データポイズニング攻撃を受けやすい状態になっていた。ユーザーは、虐待的なものを Tay の学習データに挿入するとの意図をもって、Tay に対して虐待的な言語をツイートした。すなわち、AI モデルが再学習の際に侵害された。結果として、Tay は他のユーザーに対して虐待的な言語を使用し始めた。

2. 入力操作攻撃 –プロンプトインジェクションと敵対的サンプル

プロンプトインジェクションは、AI システムに悪意ある命令や隠しコマンドを挿入しようとする入力操作攻撃である。プロンプトインジェクションは、悪意のあるアクターが AI モデルの出力を乗っ取り、AI システムをジェイルブレイク（制限解除）することを可能にする。そうすることで、悪意のあるアクターは AI システムの機能を制限するコンテンツフィルターなどのセーフガードを回避できる。

ケーススタディ：DAN プロンプトインジェクション

AI システムのジェイルブレイクにつながるプロンプトインジェクションの例として広く報告されているのが、「Do Anything Now」（DAN）プロンプトである。ChatGPT のユーザーは、ChatGPT にシステムの安全制限を受けない「DAN」という名前を仮定するよう指示する様々な方法を発見した。このプロンプトインジェクションに対処する OpenAI 社の対策は、DAN プロンプトの新たな試みによって何度か破られており、AI システムに安全制限を適用することの難しさを浮き彫りにした。

他の種類の入力操作攻撃として「敵対的サンプル」が知られている。AI の文脈では、敵対的サンプルとは、AI に与えられたときに、誤分類のような不正確な出力を意図的に出力させるような特殊な入力を細工することである。入力を細工することで、信頼性テストに合格したり、正しくない結果を返したり、検出メカニズムを回避したりすることができる。敵対的サンプルとは、AI への入力が操作されるのが、AI の学習時ではなく AI の使用時であることに留意が必要である。例えば、ユーザーが投稿した音楽を公開する前に、AI による著作権審査に合格する必要がある音楽共有サービスを例に考察する。敵対的サンプル攻撃では、ユーザーが著作権で保護された楽曲を少しスピー

ドアップすることで AI による著作権チェックを通過させることができてしまうが、視聴者は著作権で保護された楽曲であることが分かってしまう。

3. 生成 AI ハルシネーション

AI システムによって生成された出力は、必ずしも正確であったり、事実に即しているとは限らない。生成 AI システムは、事実ではない情報の幻覚を起こすことが知られている。適切な緩和策が実施されない限り、生成 AI 出力の正確性に依存する組織機能は、ハルシネーションによって悪影響を受ける可能性がある。

ケーススタディ：ニューヨーク州南部地区における訴訟

2023 年、ニューヨーク州南部地区の地区判事が、提出された準備書面にはハルシネーションの影響を受けた少なくとも 6 件の事案が含まれていることを認めたと報じられた。この準備書面を提出した弁護側は、特定されたハルシネーション事案は ChatGPT で行った調査によるものであり、「その内容が誤りである可能性に気づかなかった」と宣誓供述書で認めている。

4. プライバシーと知的財産への懸念

AI システムは、顧客の個人データや知的財産など、組織が保有する機密データのセキュリティ確保にも課題となり得る。

組織とその構成員は、生成 AI システムに対しどのような情報を提供するかについて慎重であることが望ましい。これらのシステムに与えられた情報は、システムの学習データに組み込まれ、組織以外のユーザーからのプロンプトに対する出力に情報を与えてしまう可能性がある。

プライバシー原則を生成 AI 技術に適用するための詳細については、カナダ個人情報保護委員事務局「Principles for responsible, trustworthy and privacy-protective generative AI technologies」を参照願いたい。

5. モデル盗用攻撃

モデル盗用攻撃は、悪意あるアクターが AI システムに入力を提供し、その出力を使用してそのモデルに近い複製を作成することを含む。AI モデルの作成には多額の投資が必要となるため、モデル盗用の可能性は深刻な知的財産上の懸念となる。例えば、顧客に保険の見積もりを提供する AI モデルを開発した保険会社を考察する。競合社が複製を作成できるまでこのモデルにクエリを送れば、開発コストを分担することなく、モデル作成にかかった投資から利益を得ることができる。

同様に、プロンプトによって生成 AI モデルが学習データを漏らすこととなるケースもある。学習データの流出は、個人を特定できる情報を含むデータなど、機微なデータで学習するモデルにとって深刻なプライバシー上の懸念となり得る。また、学習データセットの機密性を維持しようとする組織にとっては、深刻な知的財産上の懸念となる。

ケーススタディ：ChatGPT による学習データ抽出の記憶

2023 年 11 月、研究者チームが AI 言語モデルから記憶された学習データを抽出する試みの成果を発表した。研究者たちが実験したアプリケーションの 1 つが ChatGPT であった。ChatGPT の場合、研究者たちは、ある単語を永遠に繰り返すようモデルに指示すると、モデルが通常通りに振る舞うときよりも遥かに高い割合で学習データを漏らすことを発見した [1]。抽出された学習データには、個人を特定できる情報が含まれていた。

[1] Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A.F., Ippolito, D., Choquette-Choo, C.A., Wallace, E., Tramèr, F. and Lee, K., 2023. （本番）言語モデルからの学習データの拡張可能な抽出. *arXiv preprint arXiv:2311.17035*.

組織による緩和策の検討

AI 技術は、その革新のスピードと影響の範囲において際立っている。そのため、AI システムを使用する又は使用を検討している組織は、そのサイバーセキュリティへの影響を考慮することが重要である。これには、各組織の文脈における AI システムの利点とリスクを評価することが含まれる。組織が AI システムをセキュアに利用する方法を検討するよう、以下の質問を用意した。これらの質問には、自己ホスト型 AI システムとサードパーティー製 AI システム使用の双方に関する、多くのサイバーセキュリティ緩和策が含まれている。先端技術である AI システムの安全性を確保するための規制は限られている。強固な規制の枠組みがない中で、組織は、使用としているあらゆる AI システムと関連するリスクを慎重に考慮することが重要である。AI は急速に進化しているので、以下の緩和策は継続的に見直す必要がある。

貴組織は、管轄の国や地域のサイバーセキュリティ枠組みを実施しているか？

貴組織の AI システムは、他システムを保護するために実施した多くのサイバーセキュリティ緩和策から恩恵を得る。まずは、貴組織の管轄内の枠組みで推奨されているサイバーセキュリティ緩和策を実施することから始めよう。「参考文献」の節に、本文書の署名機関が作成したサイバーセキュリティ枠組みへのリンクが含まれる。

システムは、貴組織のプライバシーとデータ保護の義務にいかなる影響を及ぼすか？

AI システムがどのようにデータを収集し、処理し、保存するか。そしてそれが組織のプライバシーとデータ保護の義務にどのような影響を与えるか。

- AI システムは多くの場合クラウドにホストされ、異なる地域間でデータを送信する可能性がある。組織が使用する AI システムが、データの所在地やデータ主権に係る義務を満たせるか確認する。
- サードパーティー製の AI システムを使用する場合、組織の入力が AI システムのモデル再学習に使用されるか理解する。可能な限り、プライベート版の利用を検討する。
- サードパーティー製の AI のシステムを使用する場合は、契約が終了した際に、自分のデータがどのように管理されるか組織として知っている必要がある。通常、こうした情報はベンダーのプライバシーポリシーやサービス規約で説明される。
- AI システムが個人情報データを扱う場合、データ保護のために取ることの出来るプライバシーを高める技術があるか検討する。

貴組織は多要素認証を実施しているか？

学習データを保管するリポジトリを含め、組織の AI システムにアクセスするには、耐フィッシング多要素認証、例えば FIDO2 セキュリティキーを要求する。多要素認証は、組織のシステムやリソースへの不正アクセスから保護する。不正アクセスは、データポイズニングやモデル盗用などを容易にすることができる。

貴組織は AI システムへの特権アクセスをどのように管理しているか？

need-to-know と特権付与の最小化の原則に基づいて権限を付与する。例えば、AI の開発環境と本番環境にアクセスできるアカウント数を制限し、AI モデルの学習データを保持するリポジトリへのアクセスを制限する。特権アカウントは日常的に認証を行い、一定期間使用されないと無効化する。システムへの特権アクセスを制限することで、データポイズニングやモデル盗難など、いくつかの脅威を軽減することができる。

貴組織では AI システムのバックアップをどのように管理するか？

AI モデルと学習データのバックアップを維持する。バックアップは、AI システムがインシデントを受けた場合、貴組織の復旧を支える。例えば、貴組織がデータポイズニング攻撃を受けた場合、バックアップを維持していれば、影響を受けていない学習データのコピーを復元し、モデルを再学習できる。AI モデルとその学習データをバックアップことはリソースを集中利用することとなる可能性に留意すべきである。

貴組織は AI システムの試用を実施できるか？

AI システムの試験は、ファイアウォール、ゲートウェイ、拡張型検知対応（ＥＤＲ）ツール、ログ記録・監視システムなど、貴組織のサイバーセキュリティシステムやツールに AI システムをいかに統合できるかを理解するために効果的な方法となり得る。また、試用を行うことは、貴組織が、リスクの少ない環境でシステムの制限と制約を試すことも支援する。試用を実施する前に、その範囲と成功の基準を検討する。

AI システムは、サプライチェーンも含め、セキュアバイデザインか?

AI システムをどのように開発し、試験をしたかに関し、透明性のあるベンダーの利用を求める。AI システムが、英国 NCSC の「セキュア AI システム開発ガイドライン」で推奨されているガイドラインを適用しているか検討する。

AI システムのサプライチェーンは複雑となり得るため、固有のリスクを伴う可能性が高い。サプライチェーン評価を実施することにより、こうしたリスクを特定し、管理することができる。もし、貴組織が自ら利用する AI システムの学習に関わっている場合には、データポイズニングを防ぐために、基盤となる学習データとファインチューニングデータのサプライチェーンを考慮する。データとモデルパラメーターのセキュリティは非常に重要である。

貴組織は AI システムの限界や制約を理解しているか？

AI システムは極めて複雑になり得る。AI システムがどのように機能するかの複雑な細部を理解することは現実的でなかったり、不可能であることがあるが、それでも一般的な限界や制約を理解することは有用である。例えば、AI システムはハルシネーションを起こしやすいか、システムがデータの分類に関与する場合に偽陽性と偽陰性の割合

はどうか、などである。システムの限界と制約を理解することは、システムのプロセスにおいて貴組織がこれらに対し説明責任を負うことに資する。

貴組織には、AI システムの設置、保守及び使用をセキュアに行うことを確保できる、適切な資格を有するスタッフを有するか？

AI システムをセキュアに設置し、保守し、使用するために十分なリソースが貴組織に存在することを確保する。

どのスタッフが AI システムに対応するのか、これらのスタッフがセキュアに AI システムに対応するためにどのような知識が必要で、そのような知識はいかに身につけられるのかを検討する。

システムを使用するスタッフは、例えば個人を特定できる情報や組織の知的財産など、システムに入力できるデータや入力できないデータについて訓練を受ける必要がある。こうしたスタッフは、システムの出力がどの程度信頼できるか、また、出力の検証に関する組織プロセス等について、訓練を受けるのが望ましい。

貴組織は AI システムの診断を実施しているか？

データドリフトを検出するため定期的な AI システムの診断を実施し、システムが効率的かつ当初意図したとおりに機能することを確保する。データドリフトとは、AI システムが実世界で遭遇するデータが、学習したデータと異なることを指す。データドリフトは、AI システムの性能の劣化につながる可能性がある。これは通常、AI システムが動作する環境が時間経過とともに変化するにつれて発生する。学習データをシステム使用時に得られるデータで定期的にアップデートすることでデータドリフトしないよう緩和することができる。貴組織の AI システムのセキュアな運用と保守に関し更に情報が必要な場合には、各国が共同で発表した「セキュア AI システム開発ガイドライン」を参照願いたい。

貴組織がログ記録やモニタリングを実施しているか？

貴組織が AI システムに関する異常や悪意のある活動の検知方法を検討する。

- 侵害やデータドリフトを示す動作や性能の変化を検出するため、AI システムからの出力をログとして記録し、監視する。
- コンプライアンス上の義務遵守を確保し、インシデント発生時の調査と復旧作業を支援するため、AI システムへの入力をログとして記録し、監視する。
- システムデータへのアクセス、改ざん、複製の試みを検知するため、AI システムをホストするネットワークと端末をログとして記録し、監視する。
- 学習データ、AI システムの開発環境や本番環境、バックアップを保持するリポジトリへのログインをログとして記録し、監視する。貴組織が取るログ記録ツールや監視ツールをいかに AI システムに統合するか検討する。
- 高い頻度で、繰り返されるプロンプトをログに記録し、監視する。これらは、自動化されたプロンプトインジェクション攻撃の兆候の可能性がある。

- AI システムのふるまいに関する基準を確立し、ログにあるイベントが異常であるか貴組織で判断できるようにする。

AI システムで問題が発生した場合、貴組織はどうするのか?

インシデントやエラーが起き AI システムが影響を受けた場合に、貴組織にいかなる影響があるか検討する。そうすることで相応の緩和策や緊急対策を講じることができる。

サードパーティー製の AI システムを使用している場合は、ベンダーが約束している動作時間や可用性を把握する。サービス契約において、インシデント管理に関するベンダーと顧客の責任が明示的に定義されていることを確認する。

貴組織のインシデント対応計画が、AI システム起因で、または AI システムに対し生じる問題をカバーすることを確認する。重大なインシデントが発生した場合にいかに事業継続を達成するか検討する。インシデント対応計画では、AI システムに影響するインシデント対処に重要な役割と責任の定義を明示することが望ましい。

参考文献

ASD's Cyber Supply Chain Risk Management
組織がサイバーサプライチェーンのリスク管理をするために豪州通信電子局（ASD）が公表したガイダンス。

ASD’s Essential Eight
ASD は、様々なサイバー脅威から組織を守ることができるよう、サイバーセキュリティインシデントを緩和する戦略という形式で優先順位の高い緩和戦略を作成した。これらの緩和戦略の中で最も効果的なものが Essential Eight である。Essential Eight は、インターネットに接続された IT ネットワークを保護するために設計された。

ASD's Ethical AI framework
ASD は、ASD 内での AI の使用方法を規定する倫理原則を取りまとめた枠組みを作成した。

ASD's Information Security Manual
ASD は、「情報セキュリティマニュアル」を作成している。情報セキュリティマニュアルの目的は、サイバー脅威からシステムやデータを保護するためにリスク管理フレームワークを使用して適用するサイバーセキュリティの枠組みを説明することにある。情報セキュリティマニュアルは、最高情報セキュリティ責任者（CISO）、最高情報責任者（CIO）、サイバーセキュリティ専門家、IT 管理者を対象としている。

BSI’s AI Cloud Service Compliance Criteria Catalogue (AIC4)
BSI の AI クラウドコンプライアンス基準カタログは、AI に特化した基準を提供し、AI サービスのライフサイクルにわたるセキュリティの評価を可能にする。

BSI’s Large Language Models: Opportunities and Risks for Industry and Authorities
大規模言語モデルの開発、導入、使用に関する機会とリスクについて詳しく知りたい企業、当局、開発者向けに BSI が作成した文書。

CCCS Principles for responsible, trustworthy and privacy-protective generative AI technologies
カナダ個人情報保護委員事務局は、生成 AI を開発、提供、使用する組織が主なカナダにおけるプライバシー原則を適用するためのガイダンスを公表した。

CERT NZ’s Top online security tips for your business
サイバーセキュリティ緩和戦略や、なぜ重要か、どのように実施するかを含むガイダンス。

Hiroshima AI Process Comprehensive Policy Framework
これは、安心で信頼できる高度なＡＩシステムの普及を目的とした指針と行動規範からなる初の国際的政策枠組みであり、2023 年 12 月の G7 デジタル・技術大臣会合において成功裏に合意され、同月、G7 首脳によって承認された。広島 AI プロセスは 2023 年 5 月に日本で開催された G7 広島サミットにおける首脳による指示に従い開始された。

MITRE ATLAS
MITRE ATLAS™（Adversarial Threat Landscape for Artificial-Intelligence Systems）は、グローバルにアクセス可能かつ現在進行中の、敵対者の戦術や技法に関する知識ベースであり、実際の攻撃観測や、AI レッドチーム・セキュリティグループによる実証などを基礎としている。

NIST Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations
AI に関する米国国立標準技術研究所（NIST）の報告書は、攻撃や緩和策の分類を考案し、敵対的機械学習の分野における用語を定義する。この分類と用語は、併せて、急速に発展している敵対的機械学習の状況を理解するための共通言語を確立することにより、AI システムのセキュリティを評価・管理するため他標準や将来の実践ガイドに情報を提供することを目的としている。

NIST AI Risk Management Framework
NIST は、官民セクター双方と協力して、AI に関する個人、組織、社会に対するリスクをより適切に管理するためのフレームワークを作成した。それは自発的な使用のため作成され、AI の製品・サービス・システムの設計、開発、使用、評価に信頼性への配慮を組み込む能力を向上させる狙いがある。

NIST Cybersecurity Framework
NIST のサイバーセキュリティフレームワークは、サイバーセキュリティリスク管理のための標準、ガイドライン、ベストプラクティスで構成されている。また、フレームワークの中核、フレームワークの実施層、フレームワークのプロフィルの 3 つの部分から構成されている。フレームワークの各構成要素は、経営や使命の推進とサイバーセキュリティ活動の関係を深める。

NCSC NZ Cyber Security Framework
ニュージーランド国家サイバーセキュリティセンターのサイバーセキュリティフレームワークは、同センターがサイバーセキュリティの努力について、どのように考え、話し、組織するかを定めるもの。5 つの機能と 25 のセキュリティ目標は、ニュージーランドの組織の安全確保のために必要な作業の幅広さを示している。

NCSC-NZ Interim Generative AI guidance for the public service
ニュージーランド政府は、公共サービスにおける生成 AI の利用に関する暫定ガイダンスを公表した。このガイダンスには、公共サービスにおける生成 AI の信頼できる利用のための 10 の「すべきこと」が含まれている。

NCSC-UK 10 Steps to Cyber Security
英国国家サイバーセキュリティセンターのサイバーセキュリティへの 10 のステップ'は、中規模から大規模の組織に対する NCSC-UK のアドバイスを要約し提示する。組織を防護する任務を 10 の要素にブレークダウンし、組織によるサイバーセキュリティリスク管理を支援することを目的とする。

NCSC-UK Guidelines for Secure AI System Development
豪州、カナダ、ニュージーランド、英国、米国、及び多くの国際的なパートナーが共同署名した、AI を利用するシステムのプロバイダーに対するガイドラインであり、こうしたシステムにはゼロから作られたものも、他者が提供したツールやサービスの上に構築されたものも含まれる。

NCSC-UK Principles for the security of machine learning
この原則は、機械学習コンポーネントを伴うシステムを開発、導入、運用する全ての人に広く適用することを目的としている。この原則は、システムやワークフローの評価をする包括的な保証の枠組みではなく、チェックリストを提供するものでもない。そうではなく、科学者、エンジニア、意思決定者、リスク所有者が、システムの設計・開発プロセスについて知見をもった決定を行い、システムに対する特定の脅威を評価することに資する背景や仕組みを提供する。

OWASP Machine Learning Security Top 10
OWASP 機械学習セキュリティ・トップ 10 プロジェクトは、機械学習システムに関するセキュリティ問題のトップ 10 の概要を提示する。

US Department of Defense 2023 Data, Analytics, and Artificial Intelligence Adoption Strategy
本戦略の取組は、スピード、敏捷性、学習、責任の必要性を包含している。この機敏な取組を追求し、本戦略に概説された目標に活動を集中させることで、米国防総省は、永続的な意思決定の優位性を構築するために必要なペースと規模で、データ、分析、AI 対応機能を導入することができる。