AUTOCRYPT

Declaration on Cybersecurity

サイバーセキュリティ宣言

2026 年 10 月 5 日

序文

AI が自らハッキングを行いはじめました。ますますデジタル化が加速する現代社会は、深刻なサイバーセキュリティの脅威に直面しています。AI 時代のサイバー脅威をすべて正確に予測することは困難です。だからこそ、明確で一貫した基準を定め、それを個々の状況に当てはめながら課題を解決していくことが重要です。

こうした考えのもとで、ペンタセキュリティとオートクリプトの暗号学、ソフトウェア工学、フィジカル AI などの専門家が集まり、前述の基準に関する議論を重ねました。より包括的で効果的な議論と協力を生み出す出発点となることを願い、その内容を共有いたします。

2026 年 10 月 5 日

I. サイバーセキュリティ宣言

Key Element. サイバーセキュリティは社会の基本構成要素です。

私たちはデジタル化された時代を生きています。すべてのデジタルシステムにおいて、サイバーセキュリティは必ず保障されなければなりません。

Comprehensiveness. サイバー脅威に対する防衛対策は隙間なく備えられなければなりません。

「攻撃コストが利益を上回れば攻撃されない」というサイバーセキュリティの前提は、AI が攻撃を仕掛ける時代には通用しません。すべてのデジタルシステムが AI による攻撃の対象であり、わずかな脆弱性であっても危険です。したがって、あらゆるサイバーセキュリティ措置を講じる必要があります。そのために「Cyber Defense 3 原則」を以下に提示します。

Security by Design. サイバーセキュリティを優先的に考慮して設計されなければなりません。

一部の重要な IT システムや自動車を含む一部のフィジカル AI では、設計段階からサイバーセキュリティが適用されていますが、他のあらゆるデジタルシステムにおいても同様であるべきです。サイバーセキュリティの概念がシステムの核心となる機能設計思想に組み込まれていなければなりません。

> **適用例—自動車・フィジカル AI** 自動車分野では UN-R155 と ISO/SAE 21434 に基づき、設計段階からサイバーセキュリティを組み込む枠組みが、型式認証の要件として定着しています。この枠組みは、他のデジタルシステムへと展開していくうえでの基準となります。

II. Cyber Defense 3 原則

1. Observability — サイバーセキュリティ情報の測定

すべてのデジタルシステムには、サイバーセキュリティに関連するあらゆるデータを測定・分析できる技術的手段が用意されていなければなりません。この手段はデジタルシステムの OS およびそれと相互作用する AI から独立して実装・運用される必要があります。

2. Cross-Layer Defense — すべてのアタックサーフェスに対する防衛体系とそれらの連携

デジタルシステムの構成要素間の接点は、サイバーセキュリティの観点から脆弱になりがちです。これらが連結されると、アタックサーフェス（攻撃対象領域）が形成されます。デジタルシステムの発達にともなって接点が増加するほどアタックサーフェスも拡大します。アタックサーフェスは複数の層（Layer）にわたって存在するため、層ごとに適切なサイバーセキュリティ体系を構築し、それらが水平的・垂直的に連携しなければなりません。

3. Zero Damage Design — サイバー攻撃の無効化

サイバー攻撃の最終目的は重要な情報資産の窃取、偽造・改ざん、システムの無力化、あるいは制御権の掌握です。したがって、これらの目的の達成を阻む対策を講じる必要があります。

- すべての情報資産の暗号化と安全な鍵管理
- 中核機能のレジリエンス（回復力）と体系的な復旧の確保
- 被害が波及しない Compartmental Design（区画化設計）
- 中核機能に対する Fail-Safe、Minimal Risk Maneuver、非常停止装置（Red Button）の確保

III. 安全な AI のためのサイバーセキュリティ提案

AI の潜在的リスクに対する懸念が高まる中、安全な AI を実現するための方策に関する議論や研究が活発に行われています。しかし、こうした取り組みにもかかわらず、安全基準が AI によって回避されるケースや、安全基準そのものに不備がある、あるいは悪意のある AI によって意図的に欠落させられているケースも想定しておく必要があります。そこで、AI のサイバーセキュリティリスクを外部から制御する方法として以下を提案します。

1.Trustworthy AI

AI は検証された方式で動作しなければなりません。AI に対して「許容される行動」と「許容されない行動」を、厳格かつ矛盾なく定義する必要があります。定義されていない行動は「許容されない行動」とみなします。AI のアプローチや行動は客観的に観察・記録・説明可能でなければなりません。

2.Controllable AI

Trustworthy AI を確保するため、AI の行動に対する外部からの統制権が必要です。AI の行動に関する最終的な権限は人間が持ち、すべての行動はいつでも拒否・停止できなければなりません。また、AI が稼働する物理的環境に対する人間の完全な統制権も保障される必要があります。

3. Accountable AI

AI Agent には責任主体が必ず定義されていなければなりません。問題が発生した場合、責任主体は速やかに通知し、Agent を修正する必要があります。すべての AI Agent には作成者、所有者、目的、終了条件が明記されていなければなりません。

IV. 未来に向けて

AI が示した驚異的な可能性は、大きな期待と懸念を同時に引き起こしています。その懸念の中で代表的なのがサイバーセキュリティに関するもので、急速に現実化・拡散しつつあります。近年、報じられるさまざまなハッキング事件では AI の関与が疑われています。果たして AI は脅威なのでしょうか？

AI は人間が蓄積してきた数学的・工学的知識をベースに人間によって生み出されたものであり、その仕組みや原理自体はよく知られています。私たちがまだわからないのは、AI が社会の他のシステムとどのように相互作用し、どのような結果をもたらすかということです。それが何であれ、AI を人類にとって害ではなく助けとなる方向へ導いていかなければなりません。サイバーセキュリティはこの方向性を決定づける最も重要な要素です。適切な技術の開発、関連制度の整備、そして社会全体の理解と合意形成に取り組む時が来ました。

イ・ソグ
AUTOCRYPT 会長/創業者、ソフトウェア工学、M.E.

キム・ドクス
AUTOCRYPT CEO/共同創業者、暗号学、M.E.

シム・サンギュ
AUTOCRYPT CTO/共同創業者、暗号学、Ph.D.

チョン・ソンギュン
AUTOCRYPT エンジニアリング本部長、情報セキュリティ、M.E.

チョ・ソンウ
AUTOCRYPT AI モビリティ政策研究所長、自動車 AI、技術士/Ph.D.

キム・テギュン
ペンタセキュリティ CEO、海軍情報通信士官

チョン・テジュン
ペンタセキュリティ企画室長、セキュリティソフトウェア戦略

イ・ジュンヨン
ペンタセキュリティ研究所長、セキュリティソフトウェア開発

付録：用語の定義

デジタルシステム：ソフトウェアによって動作し、データを処理したり、他のシステムと連携したりして機能を果たすシステム

フィジカル AI（Physical AI）：AI が物理的環境を認識・判断し、その結果に基づいて現実世界で直接動作するシステム

Attack Surface：攻撃者がシステムに侵入したり、機能に影響を与えたり、データを持ち出したりできる箇所の集合。外部との連結箇所と構成要素間の接点の両方を含む

Resilience：攻撃や障害に備え、その状況下でも中核機能を維持し、迅速に復旧・適応する能力

Compartmental Design：システムを隔離された区画に分け、区画間の接続を制御された経路に限定することで、ある区画の侵害が他の区画へ波及しないようにする設計手法

Fail-Safe：故障や侵害が発生した際に被害が生じないように、システムを定められた安全な状態へ移行させる設計原則

Minimal Risk Maneuver：通常の動作を続けられない際、システムが自らリスクを最小化した状態へ移行する手順

Red Button：人が一度の操作で AI またはシステムの動作を中断させ、安全な状態で停止させる非常停止手段。物理的な装置とソフトウェア上の手順の両方を含む

AI Agent：与えられた目標を達成するために作業手順を自ら判断し、ツールや他のシステムを用いて行動する AI

AUTOCRYPT

AUTOCRYPT 株式会社 115 Yeouigongwon-ro, Yeongdeungpo-gu, Seoul, 07241 · www.autocrypt.io

© 2026 AUTOCRYPT Co., Ltd. All rights reserved.