Acompany
アカンパニー

Confidential AI : Confidential Computing で実現する「秘密を守れるAI」

Date：2026 年 9 月 1 日

Contact：marketing@acompany-ac.com

Confidential AI：Confidential Computingで実現する「秘密を守れるAI」

株式会社Acompany

／ ABSTRACT

本稿は、「秘密を守れる AI」を具現化する **Confidential AI** というパラダイムについて、その要件、背景技術、期待される価値、および今後の展望を整理する。生成 AI の産業利用が急速に進む中、利用者は機密データを自らの統制下に留めたい一方、モデル提供者はモデルの重みや知的財産を自らの管理下に留めたいという、対称的な緊張関係が生じる。この状況で「秘密を守れる AI」を実現するには、AI の出力や行動を適切に制御する振る舞いの信頼性に加えて、計算中のデータやモデルを保護し、実行環境を検証できる環境の信頼性が必要となる。

本稿では、Trusted Execution Environment（TEE）と Remote Attestation（RA）を中核とする **Confidential Computing（CC）**を基盤として、実行環境と振る舞いの両面から「秘密を守れる AI」を具現化する技術・運用パラダイムを **Confidential AI** と定義する。さらに、CC 対応計算基盤のスケールアップに伴う適用可能性の拡大と、エージェント協働やフィジカル AI など、Confidential AI の新たな応用領域への展開を展望する。

Date：2026 年 9 月 1 日

Contact：marketing@acompany-ac.com

CHAPTER 01

AI ネイティブな社会に求められる「秘密を守れる AI」

生成 AI の業務利用を検討する組織は、次の問いに直面する。

「このデータを、組織外の AI サービスに送信してよいのか」

融資審査の資料、患者の診療情報、製品の設計図などを、高性能なフロンティアモデルで処理したい。しかし、外部の AI サービスを利用するには、機密データを自組織の統制が及ばない環境へ送信する必要がある。一方、オープンウェイトモデルの自組織での運用には、モデル性能、計算資源、運用負荷、継続的な更新といった制約がある。このため、利用者は、データへの統制の維持と、高性能な AI の利用との間で難しい選択を迫られる。

同じ問題は、モデル提供者の側にも存在する。モデルの重み（パラメータ）は、研究開発投資、学習データ、設計上のノウハウを反映した重要な知的財産である。これをオンプレミス等の顧客環境へ配備すれば、重みの読み出し、複製、不正利用といったリスクが生じる。そのため、モデル提供者はモデルを自らの管理下に置き、API を介して提供することが多い。

すなわち、**利用者はデータを手元に留めたい一方、提供者はモデルを自らの管理下に留めたい。**生成 AI の産業利用には、この対称的な緊張関係が存在する。

利用者とモデル提供者の双方にとって、「秘密を守れる AI」はどのように実現できるのだろうか。

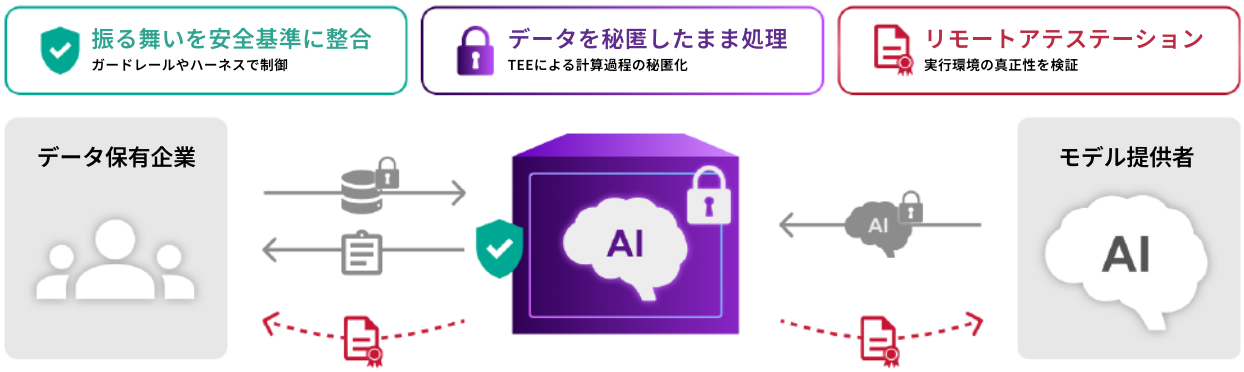

図 1　Confidential AI の概念。CC による実行環境の保護・検証を中核に、AI セキュリティ技術による振る舞いの制御を組み合わせることで、「秘密を守れる AI」を具現化する。

これまで、有害な出力の抑制、プロンプトインジェクションへの対処、エージェントの権限制御など、さまざまな AI セキュリティ技術が研究・開発されてきた。これらは、主として **AI が何を出力し、どのように振る舞うか**を制御する技術であり、安全な AI システムを構築する上で不可欠である。

しかし、これらの技術は、**AI がどのような環境で実行され、その環境において誰が入力、出力、モデルへアクセスできるのか**という問題には直接答えない。モデルの振る舞いが適切に制御されていても、実行環境を管理する主体が機密情報へアクセスできるのであれば、機密性の高い用途に利用することは難しい。

1.1 Confidential Computing が導く Confidential AI というパラダイム

この問いに答える計算パラダイムとして、Confidential Computing（機密コンピューティング、CC）が近年注目されている。

CC は、計算中のコードやデータ等を隔離する Trusted Execution Environment（TEE）[9] と、その実行環境を暗号学的に検証する Remote Attestation（RA）[4] で構成される。具体的には、TEE による秘密を守れる実行環境でデータを処理することで、OS やハイパーバイザ、クラウド管理者など、TEE の保護境界の外にいる特権主体からデータを秘匿する。RA は、実行環境が本当に秘密を守れる環境であるかを検証する。

本稿では、CC を中核として AI セキュリティ技術を組み合わせ、実行環境と振る舞いの両面から「秘密を守れる AI」を具現化する技術・運用パラダイムを、**Confidential AI** と定義する（図 1）。Confidential AI では、利用者は、モデル提供者やインフラ事業者へデータを秘匿し、機密性を保ったまま、高性能なモデルを利用できる。一方、モデル提供者は、自らの管理外にある環境へモデルを配備する場合でも、重みへの直接的なアクセスを制限できる。さらに、ガードレールやハーネスなどを組み合わせることで、悪意ある入力、有害な出力、エージェントの権限逸脱といった振る舞い上のリスクにも対処する。すなわち、Confidential AI によれば、利用者とモデル提供者がそれぞれの秘密への統制を維持しながら、AI を信頼できる形で提供・運用することが可能となる。

1.2 Confidential AI への期待

Confidential AI の意義は、現在の生成 AI サービスにおけるデータやモデルの保護だけに留まらない。近年、CC に対応したハードウェアの進展により、CC の保護範囲が多数の CPU や GPU を含む大規模な計算基盤へ拡張され、Confidential AI をフロンティアモデルへ適用する道が開かれようとしている。その先には、未修正の脆弱性を秘匿したまま解析する AI [3, 7]、互いの実行環境を検証しながら協働するエージェント群、現場のデータや知的財産を保護するフィジカル AI 基盤などが見えてくる。

これらのフロンティアの実現には、大規模な計算資源、RA の検証基盤、鍵管理、セキュアな通信、監視・運用を一体として提供する CC に対応した AI データセンターの整備が重要となる。

■本稿の目的と構成　本稿は、Confidential AI の意義を明確にすることを目的とする。2 章で秘密を守れる AI の要件を整理し、3 章で CC の仕組みを解説する。4 章にて、Confidential AI の定義と提供する価値について概観する。5 章にて、Confidential AI が切り拓くフロンティアを展望する。

CHAPTER 02

「秘密を守れる AI」に求められる信頼性

1 章では、利用者のデータと提供者のモデルを保護しながら、AI を信頼できる形で提供・運用するという「秘密を守れる AI」の考え方を示した。本章では、その実現に求められる要件を整理し、既存の AI セキュリティ技術がどの要件を担うのか、また、実行環境の面でどのような信頼性が必要となるのかを明らかにする。

2.1 二つの信頼性

「秘密を守れる AI」を実現するには、AI が扱うデータやモデルを秘匿するだけでは十分ではない。AI の入力、出力、行動を適切に制御するとともに、それらが処理される実行環境を保護・検証できる必要がある。本章では、これらの要件を振る舞いの信頼性と環境の信頼性の二つに整理する。

■振る舞いの信頼性　AI がどのような出力を生成し、どのような行動を取るかに関する信頼性である。主に、次の二つの要件から構成される。

- **入力・出力の制御：**悪意ある入力（プロンプトインジェクション等）や有害な出力に対処し、モデルの振る舞いを定められた方針や利用目的に沿うよう制御する。モデルの振る舞いが秘密を漏らさない。
- **権限と行動の統制：**AI エージェントが利用できるツール、アクセス可能な情報、操作対象、予算、実行回数などを制限し、権限の逸脱や意図しない行動を抑制する。

■環境の信頼性　AI がどのような環境で実行され、その環境において秘密がどのように扱われるかに関する信頼性である。主に、次の二つの要件から構成される。

- **計算中の秘密の保護：**入力、出力、モデルの重み、認証情報、メモリ、推論中の内部状態などを、データオーナー以外（OS、ハイパーバイザ、インフラ管理者など）による不正な閲覧や改ざんから保護する。
- **実行環境の検証：**期待したハードウェア、ソフトウェア、モデル、ガードレールなどが、実際の環境で使用されていることを外部から検証可能にする。

二つの信頼性と、それぞれを構成する要件および主な技術の関係を、表 1 に示す。実際の AI システムには、これらに加えて、可用性、正確性、監査可能性、データガバナンス、物理的安全性なども求められる。本稿では、秘密の保護と AI の安全な振る舞いを両立するための信頼性に焦点を当てる。

表 1「秘密を守れる AI」に求められる信頼性と主な技術

信頼性	構成要件	主な対象	主に担う技術
振る舞いの信頼性	入力・出力の制御	モデルへの入力と生成される出力	アラインメント、ガードレール
振る舞いの信頼性	権限と行動の統制	ツール、操作対象、権限、予算	ハーネス、認証・認可、監視
環境の信頼性	計算中の秘密の保護	入力、出力、モデル、 認証情報、内部状態	Confidential Computing、 暗号化、アクセス制御
環境の信頼性	実行環境の検証	ハードウェア、コード、 モデル、システム構成	Remote Attestation

2.2 振る舞いの制御と AI セキュリティ

従来の AI セキュリティ技術は、前述の振る舞いの信頼性について配慮してきた。AI が作用する局面に応じて、次のように整理できる。

- **訓練時**　アラインメント（Safety Alignment、AL）は、モデルの訓練や追加学習を通じて、その振る舞いを望ましい方向へ調整する。有害な出力の抑制や、指示・方針との整合が主な対象となる。
- **推論時**　ガードレール（GR）は、運用中のモデルへの入力と出力を監視・制御する。プロンプトインジェクションやポリシーに反する入力の検知、不適切な出力の遮断などを担う。
- **エージェント実行時**　ハーネス（HN）は、AI エージェントが利用できるツール、権限、操作対象、実行回数、予算などを制限する実行枠組みである。複数のエージェントからなるシステムでは、タスクの割り当てや行動の調停を担うオーケストレーションも必要となる。

これらの技術は、訓練、推論、エージェント実行という異なる局面で、AI の出力や行動を制御する。すなわち、「秘密を守れる AI」に求められる要件のうち、主として振る舞いの制御と権限と行動の統制を担う。一方、これらは、モデルがどの実行環境で動作しているのか、その環境の管理者が計算中のデータやモデルへアクセスできるのか、宣言されたモデルやガードレールが実際に使用されているのかを保証するものではない。これは技術的な不足ではなく、各技術が対象とする役割の違いによるものである。

2.3 実行環境に求められる信頼性

機密情報を扱う AI では、モデルの振る舞いだけでなく、**計算が行われる環境そのものをどこまで信頼できるか**が重要となる。具体的には、計算中のデータやモデルへ誰がアクセスできるのか、期待したコードや構成が実際に使用されているのかを確認する必要がある。

オープンウェイトモデルを自組織の計算基盤で運用することは、この問題に対する有力な選択肢である。ただし、大規模なモデルの運用にクラウド上の GPU 基盤を利用する場合には、計算中のデータやモデルをインフラ事業者からどのように保護するかという課題が残る。重要なのは、モデルがローカルに存在するか否かだけでなく、計算基盤を誰が管理し、どの構成要素を信頼する必要があるかである。

また、Zero Data Retention（ZDR）、Data Processing Agreement（DPA）、内部統制、第三者監査などは、責任の所在を明確にし、組織的・法的な保証を与える上で重要である。ただし、これらは、運用主体によるデータへのアクセスを技術的に制限したり、特定の実行時点における環境の状態を検証したりするものではない。

したがって、「秘密を守れる AI」には、振る舞いの制御や運用上の統制に加えて、**明示された脅威モデルの下で、計算中の秘密を保護し、期待した実行環境が使用されていることを検証可能にする仕組み**が必要となる。

2.4 Confidential AI が具現化する「秘密を守れる AI」

ここまでの整理から、「秘密を守れる AI」には、振る舞いの信頼性と環境の信頼性の双方が必要である。両者は代替関係ではなく、互いに補完する関係にある。

AI の入力、出力、行動が適切に制御されていても、実行環境からデータやモデルが漏洩するのであれば、秘密を守れる AI とはいえない。反対に、計算中の秘密が保護されていても、モデルが有害な出力を生成したり、エージェントが権限を逸脱したりするのであれば、AI を信頼できる形で運用することはできない。

1章では、**Confidential AI**を、Confidential Computingを中核として環境の信頼性を構成し、アラインメント、ガードレール、ハーネスなどのAIセキュリティ技術による振る舞いの信頼性を組み合わせることで、「秘密を守れるAI」を具現化する技術・運用パラダイムと定義した。実際のシステムでは、これらに加えて、鍵管理、認証・認可、監視、監査などを含む多層的な設計が必要となる。

続く 3 章では、環境の信頼性を担う中核技術として、Trusted Execution Environment と Remote Attestation の仕組み、保証範囲、および設計上の考慮事項を説明する。その上で、4 章では、Confidential AI がデータとモデルの保護、実行環境の検証、主権への寄与といった価値をどのように生み出すのかを整理する。

CHAPTER 03

Confidential Computing

2 章で確認した、環境の信頼性を技術的に担う手段の空白を埋める役割を、Confidential Computing（機密コンピューティング、CC）は担う。CC は、二つの中核的な要素から構成される。一つは、計算中のデータの機密性と完全性を保護する Trusted Execution Environment（TEE）である。もう一つは、その実行環境の真正性を検証可能にする Remote Attestation（RA）である。本章では、これらを順に説明した上で、CC の保証範囲と設計上の考慮事項を整理する。

3.1 Trusted Execution Environment

データの保護は、伝統的に「保存時（at rest）」と「転送時（in transit）」の二つの局面で語られてきた。保存時のデータはディスク暗号化によって、転送時のデータは TLS などの通信暗号化によって保護される。

しかし、データが実際の計算に供される瞬間、すなわち**使用時（in use）**には、通常、データを平文としてメモリ上に展開する必要がある。ここに、従来の暗号化では十分に保護されてこなかった空白が生じる。2 章で述べた、他者が管理するインフラ上で計算することへの不信が問題となるのは、まさにこの局面である。この空白を埋めるのが TEE である。

TEE は、CPU や GPU などのハードウェアが提供する**隔離された実行環境**である（図 2）。その内部で処理されるコードとデータを、同一マシン上で動作する他のソフトウェアから保護する。対象となる脅威モデルの範囲内では、OS、ハイパーバイザ、さらにはクラウド管理者のような特権主体であっても、TEE 内部のコードやデータを直接観測することはできない。また、コードやデータに対する不正な改ざんを検知し、改ざんされた状態で処理が継続されることを防ぐ。

図 2 Trusted Execution Environment

本稿が主に対象とするクラウド向けの TEE では、アクセス制御に加えて、メモリ暗号化や完全性保護などをハードウェアレベルで実施する。これにより、インフラを物理的に保有・運用する事業者であっても、その上で実行される計算の内容にアクセスできないという強い隔離が実現される。

ここで重要なのは、TEE が信頼を不要にする技術ではなく、**信頼しなければならない範囲を限定する技術**であるという点である。利用者は、OS、ハイパーバイザ、クラウド事業者の運用などへの信頼を前提とするのではなく、ハードウェア、ファームウェア、TEE の保護機構、アテステーション基盤などから構成されるTrusted Computing Base（TCB）に、信頼の対象を絞り込む。したがって、TEE の本質は「誰も信頼しなくてよい」ことではなく、**何を信頼の前提とするのかを明確にし、その範囲を可能な限り小さくすること**にある。

■隔離方式に関する注意　厳密には、TEE は隔離実行技術の総称であり、すべての TEE がメモリ暗号化によって隔離を実現するわけではない。アクセス制御を中心として隔離する TEE も存在する。一方、CC の文脈、とりわけ利用者がハードウェアを物理的に管理できないクラウド環境では、コールドブート攻撃などの物理的な脅威にも対処するため、メモリ暗号化に対応した TEE が重要となる。本稿で以下「TEE」と記す場合は、特に断らない限り、このような暗号化対応 TEE を念頭に置く。

■GPU における CC　NVIDIA の Hopper 世代以降の GPU に搭載されている CC では、GPU を保護された仮想マシン Confidential Virtual Machine（CVM）に排他的に割り当てられた状態になる。これにより、同じ GPU を複数の CVM で共有せず、GPU メモリ上のモデルや入力データを、ホスト OS やハイパーバイザから隔離する。CPU 側の TEE と GPU 側の保護機構を組み合わせることで、生成 AI の推論や学習を、CPU と GPU をまたぐ保護された実行環境上で実行できる。このように、大規模な AI ワークロードを隔離実行環境上で動作させることが現実的な選択肢となりつつある。

3.2 Remote Attestation

TEE によってコードとデータが保護されていても、利用者は、その環境が正規の TEE であり、期待した構成で動作していることを確認できなければならない。この検証を可能にするのが Remote Attestation（RA）である。

RA では、TEE がハードウェア、ファームウェア、セキュリティ設定、コードや構成などの状態を計測し、その結果を含むアテステーション証拠を生成する。この証拠は、ハードウェアベンダーの信頼の根へ連なる鍵によって署名または認証される。検証者は、署名と計測値をあらかじめ定めた参照値や検証ポリシーと照合することで、正規の TEE が使用され、期待したコードや構成が動作していることを遠隔から確認できる。

RA の結果は、通常、データの送信や暗号鍵の解放と結び付けて利用される。すなわち、期待した環境であることが確認された場合にのみ、入力データやモデルを TEE へ投入する。これにより、契約や運用者の説明だけに依存せず、実行時の環境を暗号学的な証拠に基づいて検証できる。

ただし、RA が保証するのは、**検証ポリシーに一致する環境やコードが実行されていること**であり、そのコードが安全であることや、望ましい結果を生成することではない。RA は「何が動いているか」を検証可能にするが、「それが適切であるか」は別途評価する必要がある。したがって、モデルやガードレールの性質を評価する技術と、その評価済みの構成が実際に動作していることを確認する RA は、補完的な役割を担う。

3.3 保証範囲と設計上の考慮事項

CC が緩和するのは、データやモデルを相手へ平文で開示しなければ計算できないという構造的な制約である。CC の主な限界とトレードオフとして次の四点を挙げる。

- **脅威モデルの範囲**：TEE は、OS、ハイパーバイザ、クラウド管理者などの特権主体からコードとデータを保護する。一方、すべてのサイドチャネル攻撃や高度な物理攻撃まで防ぐものではないため、CC は多層防御の一部として位置付ける必要がある。

- **可用性の非保証**：TEE は、コードやデータの機密性と完全性を保護するが、実行環境の可用性までは保証しない。インフラ事業者や攻撃者による処理の停止、通信の遮断、計算資源の制限などには、冗長化や障害復旧などの別の対策が必要である。

- **性能オーバーヘッド**：メモリ暗号化や CPU–GPU 間通信の保護には、一定の性能オーバーヘッドが伴う。その影響はハードウェア、システム構成、ワークロードによって異なるため、導入時には実環境での評価が必要である。

- **TCB への依存**：CC は信頼を不要にするのではなく、信頼しなければならない範囲を TCB へ限定する。したがって、ハードウェア、ファームウェア、鍵管理、アテステーション基盤などへの信頼は残る。この依存は、4 章で述べる主権の問題にも関係する。

CHAPTER 04

Confidential AI

Confidential AI は、CC を中核として AI セキュリティ技術を組み合わせ、実行環境と振る舞いの両面から「秘密を守れる AI」を具現化する技術・運用パラダイムである。

その中で CC は、TEE によって入力、出力、モデル、認証情報、中間状態などの計算中の秘密を保護し、RA によって期待した実行環境やソフトウェア構成が使用されていることを検証可能にする。一方、アラインメント、ガードレール、ハーネスなどの AI セキュリティ技術は、AI の入力、出力、行動に関するリスクへ対処する。Confidential AI は、これらを多層的に組み合わせることで成立する。

本章では、Confidential AI がもたらす価値と代表的な利用形態について整理する。具体的には、利用者のデータ保護、提供者のモデル保護、データや AI 能力に対する主権への寄与、実行環境の検証について論じる。さらに、代表的な利用形態ごとに、考慮すべき信頼性と、その達成に必要な多層防御設計について概観を示す。全体の関係を表 2 に示す。

4.1 データの保護

Confidential AI の第一の価値は、利用者が、機密データを AI ベンダーやインフラ事業者に対して秘匿し、機密性を保ったまま、高性能なモデルで処理できることである。TEE 内で入力、推論中の状態、出力を保護することにより、他者が管理する計算基盤を利用しながら、データへの直接的なアクセスを制限できる。

これは、金融、医療、公共、製造など、機密性の高いデータを扱う領域において特に重要である。CC の導入だけで法令や業界規制への準拠が成立するわけではないが、使用中のデータへのアクセスを技術的に制限することは、規制対応を支える技術的統制の一つとなり得る。また、組織として承認された安全な利用経路を提供することは、従業員が機密情報を無断で外部サービスへ入力するシャドー AI への対策にもつながる。

■事例：Apple Private Cloud Compute [1]

Apple の Private Cloud Compute は、端末上では処理できない AI ワークロードをサーバ側で実行しながら、利用者のデータを運用者から保護し、実行環境を外部から検証可能にする設計例である。

4.2 モデルの保護

第二の価値は、モデル提供者に対して、自らが直接管理しない環境へモデルを配備する場合でも、モデルの重み（パラメータ）への直接的なアクセスを制限できることである。モデルの重みには、研究開発投資、学習データ、設計上

表 2　Confidential AI がもたらす価値

観点	主な受益者	保護・統制の対象	CC の役割
データの保護	データ保有者	入力、出力、業務データ	使用中のデータをインフラ運用者などから保護
モデルの保護	モデル提供者	モデルの重み、構成、知的財産	配備先における重みへの直接アクセスを制限
主権への寄与	国家、企業、組織	データと AI 能力に対する統制	外部の計算資源やモデル利用時の統制を補強
実行環境の検証	すべての当事者	ハードウェア、コード、構成	RA で期待した環境の利用を検証可能にする

の工夫などが反映されており、提供者にとって重要な知的財産となる。

暗号化されたモデルを TEE 内部でのみ復号・実行する構成を採れば、利用者が管理するオンプレミス環境やエッジ環境へモデルを配置しながら、ホスト OS や管理者による重みの読み出しを防止できる。これにより、API としてのみ提供していたモデルを、データに近い環境へ配備するという選択肢が生まれる。

さらに、RA と鍵管理を組み合わせることで、期待したハードウェア、ソフトウェア、利用条件が確認された場合にのみモデルの復号鍵を解放できる。ただし、CC が直接防止するのは重みの読み出しであり、モデルの出力を大量に収集するクエリベースの抽出攻撃などには、別の対策が必要である。

■事例：Google Distributed Cloud [6]

Google Distributed Cloud は、生成 AI をオンプレミス環境や外部ネットワークから隔離された環境へ配備する例として位置付けられる。このような構成では、利用者のデータに対する統制と、提供者のモデルに対する保護を同時に成立させることが重要となる。

4.3 主権への寄与

データ主権や AI 主権は、法制度、インフラ、人材、サプライチェーン、運用能力などを含む広い概念であり、CC だけで実現できるものではない。CC の役割は、外部のインフラやモデルを利用する場合にも、データとモデルへのアクセスを技術的に制限し、組織が持つ統制を補強することにある。

データ主権の観点では、データの物理的な所在だけでなく、誰がその内容へアクセスできるかが重要となる。データが自国や自組織の指定した地域に保存されていても、そのインフラの運用者がデータを閲覧できるのであれば、統制は物理的な所在だけでは完結しない。CC は、データの所在や運用主体に加えて、アクセス可能性をハードウェアによって制限する手段を提供する。

AI 主権の観点では、CC は二つの方向に寄与する。モデル提供者は、知的財産を保護しながら、自らの管理外にある環境へモデルを展開できる。一方、利用者は、自らフロンティアモデルを開発できない場合でも、データへの統制を維持しながら外部の AI 能力を利用できる。

もっとも、CC は外部技術への依存を解消するものではない。信頼の対象は、クラウドの運用者から、ハードウェア、ファームウェア、アテステーション基盤などから成る TCB へと移る。CC は主権を単独で実現するのではなく、依存の範囲を限定し、その状態を検証可能にする技術として位置付けられる。

4.4 実行環境の検証

データとモデルの保護は、期待した TEE とソフトウェア構成が実際に使用されていることを確認できて初めて意味を持つ。RA は、実行環境の計測値を検証し、その結果をデータの送信や暗号鍵の解放と結び付けることで、保護されていない環境への秘密情報の投入を防ぐ。

RA によって検証できるのは、検証ポリシーに一致する環境やコードが動作していることであり、そのコードの安全性や処理結果の正しさではない。したがって、モデルやガードレールの性質を評価する技術と、評価済みの構成が実際に使用されていることを確認する RA は、補完的な役割を担う。

同じ TEE が利用者のデータと提供者のモデルを同時に保護することで、双方はそれぞれの秘密への統制を維持したまま、同じ計算を成立させられる。さらに、RA は単一の実行環境を検証するだけでなく、異なるシステムが互いの実行環境を検証し、通信やデータ共有の可否を判断するためにも利用できる。各システムがアテステーション証拠を交換し、その結果を認証された通信セッションへ結び付けることで、組織や計算基盤の境界を越えた検証可能な信頼関係を構成できる。

RA による相手の実行環境とコードの検証は、異なる環境で稼働する AI エージェント同士が秘密を保持したまま協働するための基盤の構築にも貢献すると考えられる。

4.5 Confidential AI の多層防御設計

Confidential AI は、環境の信頼性と振る舞いの信頼性を多層的に組み合わせることで、「秘密を守れる AI」を具現化する。ただし、保護すべき資産、想定される脅威、求められる統制は利用形態によって異なるため、二つの信頼性を実現するための具体的な設計も一様ではない。

代表的な利用形態として次の四つを挙げる。四つの利用形態について、環境の信頼性と振る舞いの信頼性をそれぞれどのように構成するかを整理する（表 3）。

- **Confidential Inference**：利用者の入力、推論中の状態、出力を TEE 内で保護し、AI ベンダーやインフラ事業者へ機密性を保ったまま推論を実行する。

- **Confidential Model Deployment**：暗号化されたモデルの重みを TEE 内でのみ復号・実行することで、モデル提供者の管理外にあるオンプレミス環境やエッジ環境へ、モデルを保護したまま配備する。

- **Confidential Agents**：認証情報、コンテキスト、計画、ツールの実行結果などを、保存時は鍵管理と暗号化で、使用時は TEE を用いて保護することで、エージェントの内部状態を外部の運用者やインフラ事業者から秘匿する。

- **Confidential Training**：訓練データ、モデル、勾配、中間状態を TEE 内で保護し、データを訓練環境の運用者へ開示することなく、モデルの訓練やファインチューニングを実行する。

表3が示すように、二つの信頼性が具体的に意味する内容は、利用形態によって異なる。Confidential AIの多層防御設計では、環境と振る舞いのいずれか一方だけを強化するのではなく、利用形態ごとの脅威モデルに応じて、両者を一体的に設計することが重要である。

表 3 Confidential AI の利用形態に応じた多層防御設計

利用形態	CC が担う環境の信頼性	振る舞いの信頼性	主な構成要素
Confidential Inference	入力、出力、推論中の内部状態の保護と、推論環境の検証	悪意ある入力や有害な出力の検知・制御	TEE、RA、AL、GR、アクセス制御
Confidential Model Deployment	モデルの重みや鍵の保護と、配備されたモデルおよび実行環境の検証	モデル抽出や利用条件からの逸脱の検知・抑制	TEE、RA、レート制限、抽出検知、鍵・ライセンス管理
Confidential Agents	認証情報、メモリ、計画、実行状態の保護と、ハーネスやツール構成の検証	権限の逸脱、不正なツール実行、意図しない行動の抑制	TEE、RA、HN、GR、認証・認可、監視
Confidential Training	訓練データ、モデル、勾配、中間状態の保護と、訓練環境の検証	学習処理のデータ利用方針への適合と、訓練後のモデルを通じた情報漏洩の抑制	TEE、RA、PETs、来歴管理

CHAPTER 05

展望：Confidential AI のフロンティア

4 章では、Confidential AI がもたらすデータとモデルの保護、主権への寄与、実行環境の検証、および検証可能な透明性を整理した。本章では、その先に広がる応用を展望する。以下には構想段階のものも含まれるが、フロンティアモデルの能力向上と、CC の保護範囲の拡大によって、技術的な実現可能性は高まりつつある。

5.1 ラックスケール Confidential Computing

大規模な生成 AI の学習や推論には、多数の GPU を高速なインターコネクトで接続した計算基盤が必要となる。そのため、単一または少数の GPU だけを保護できても、フロンティアモデル全体を CC の保護境界内で実行することは難しかった。

この制約は**ラックスケール Confidential Computing** の登場によって変わろうとしている。CPU、GPU、デバイスメモリ、インターコネクトを含むラック全体へ信頼境界を拡張し、大規模な AI ワークロードを一体的に保護する。その代表例である NVIDIA Vera Rubin NVL72 では、72 基の GPU、36 基の CPU、およびインターコネクトにまたがる統一的なセキュリティ領域が構成される [2]。これにより、モデル、入力データ、中間状態、出力を保護しながら、フロンティアモデルを大規模な計算基盤上で実行できる可能性が開かれる。

5.2 秘密を守れる脆弱性診断

フロンティアモデルとエージェント型の実行基盤は、複雑なソフトウェアから未知の脆弱性を発見し、修正を支援できる水準へ到達しつつある。Anthropic の Claude Mythos Preview は、主要な OS や Web ブラウザから未知の脆弱性を発見し、長年見過ごされてきた OpenBSD の脆弱性を特定したと報告されている [3]。

また、複数エージェントの構成を最適化する研究でも、Google Chrome から未知の脆弱性が発見されている [7]。OpenAI は、Patch the Planet [8] と呼ばれる、オープンソースソフトウェアのセキュリティ向上と脆弱性修正を目的とした取り組みを 2026 年 6 月に開始した。

この用途では、診断対象となるソースコードやシステム構成だけでなく、診断によって新たに生成される脆弱性情報も保護対象となる。未修正の脆弱性やエクスプロイト情報は、漏洩すれば攻撃に直接利用され得るため、一般的な業務データ以上に厳格な秘匿性が求められる。

Confidential AI は、診断対象と診断結果の双方を TEE 内に留め、モデル提供者やインフラ事業者へ開示せずに、高度な脆弱性診断能力を利用する構成を可能にする（図 3）。これは、最も秘匿すべき情報を外部モデルへ開示しなければ、高度な診断能力を利用できないという制約を緩和する。ただし、CC による環境の信頼性だけでは、モデルの高度なサイバー能力が不正に利用されることまでは防げない。

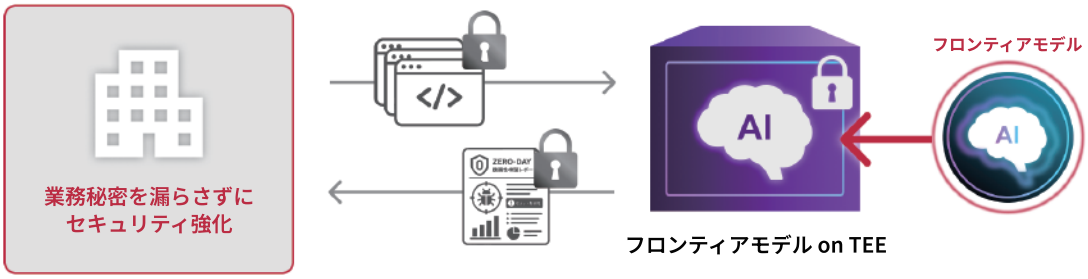

図 3　秘密を守れる脆弱性診断：Confidential AI 基盤上でフロンティアモデルを実行することで、診断対象と診断結果を保護しながら、モデル提供者やインフラ事業者へ開示することなく、高度な脆弱性診断能力を利用可能にする。

診断対象、利用者、ツール、外部通信、出力の取り扱いを制限する強固なハーネスによって振る舞いの信頼性を担保した多層防御が求められる。

5.3 秘密を守れるエージェント間協働

AI エージェントの利用は、単一のエージェントが処理を完結する形態から、異なる組織や事業者が提供するエージェント同士がタスクを委譲し、協働する形態へ拡張しつつある。Agent2Agent（A2A）プロトコルは、異なる実装や基盤で動作するエージェント間の通信と相互運用を標準化する取り組みである [10]。

通信方式が標準化されても、相手のエージェントがどのような環境とコードで動作しているかは、自動的には保証されない。また、委譲されるタスク、認証情報、内部状態、ツールの実行結果には、組織ごとの機密情報が含まれ得る。

そこで、各エージェントを TEE 内で実行し、RA によって相手の実行環境を検証した上で通信を確立する構成が考えられる。これにより、各エージェントは自らの内部情報を保護しながら、通信相手が期待したコードと構成で動作していることを確認できる。ただし、RA は通信相手の将来の振る舞いまで保証するものではない。相手へ開示する情報や付与する権限を最小限に限定し、タスクやツールの利用を継続的に監視する必要がある。

5.4 秘密を守れるフィジカル AI 基盤

Confidential AI の応用範囲は、生成 AI やソフトウェアエージェントから、ロボット、自動運転車、ドローン、センサーなどが稼働する物理世界へと拡張できる。本稿では、現場で生成されるデータや知的財産を保護しながら、フィジカル AI を開発・運用する基盤を、**秘密を守れるフィジカル AI 基盤**と呼ぶ（図 4）。

製造現場やモビリティ領域で得られる映像、センサー値、制御ログには、現場固有のノウハウや安全上の弱点が含まれる。この基盤では、これらの情報とモデル、推論中の内部状態を TEE 内で処理することで、現場の秘密を外部へ開示せずに AI を活用する。具体的には、現場データを用いた産業用 AI の学習・更新、現場の状態予測とリスク評価を行う**デジタルツイン**、その予測に基づいてロボットや車両へ高次の計画と指示を与える **AI オーケストレーション**を統合する。

ただし、TEE と RA による環境の信頼性だけでは、予測の正確性やロボットの安全な行動まで保証できない。センサーの真正性、モデルの精度、エージェントの権限制御、リアルタイム制約、物理的な安全機構などを組み合わせ、振る舞いの信頼性を含む多層防御を構成する必要がある。

フィジカル AI そのものも研究途上の技術であり、多大な投資がされている。フィジカル AI が社会実装された際に、データ主権を確保しながら活用できるようにするためにも、CC に対応したフィジカル AI 基盤の整備に向けた研究開発も同様に注力していくべき領域であると考える。

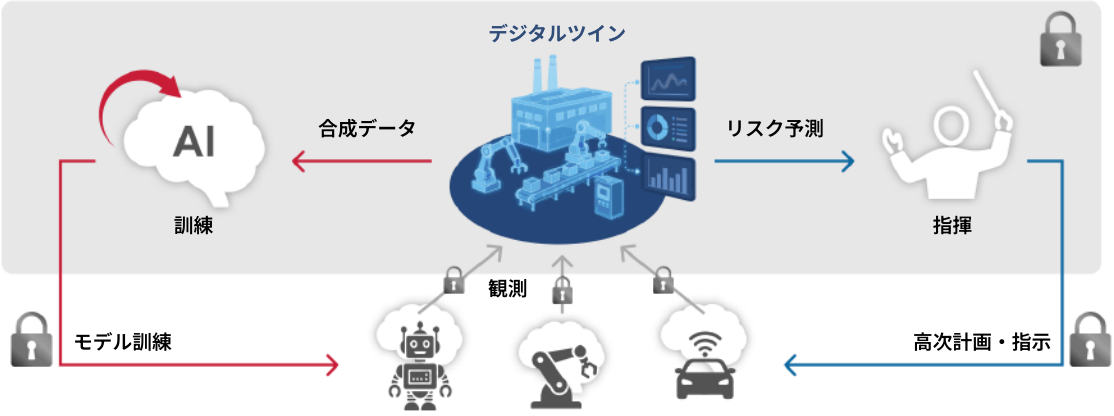

図 4　秘密を守れるフィジカル AI 基盤：Confidential AI に基づくプラットフォームにより、現場の秘密を守りながら、フィジカル AI を業務に最適化する。VLA 等のロボット基盤モデルの訓練・推論と、フィジカルエージェントのオーケストレーションによる最適配置を、データ主権を維持しながら実現する。

5.5 社会基盤としてのConfidential AI

本章で述べた脆弱性診断、エージェント間協働、フィジカル AI は、いずれも一部の組織だけに閉じた用途ではない。金融、医療、製造、公共、安全保障など、機密性の高い情報を扱う幅広い領域において、外部の高度な AI 能力を安全に利用するための共通基盤となり得る。

生成 AI が社会や産業の基盤へ組み込まれるほど、求められるのは高性能なモデルだけではなく、誰がデータやモデルへアクセスできるのかを技術的に制御し、その状態を検証できる計算環境である。Confidential AI は、CC による環境の信頼性と、AI セキュリティ技術や運用上の統制による振る舞いの信頼性を組み合わせ、利用者のデータと提供者のモデルを保護しながら、組織の境界を越えた AI 利用を可能にする。将来的に、Confidential AI が、AI 活用における標準的なセキュリティ施策や基盤を担っていくことになるだろう。

誰もが Confidential AI の恩恵を受けられる社会基盤を実現するためには、**CC に対応した AI データセンターの整備が不可欠となる**。大規模な CC 対応計算資源を共通基盤として提供することで、自ら専用インフラを保有できない組織も、データやモデルへの統制を維持しながら、高度な AI を利用できるようになる。また、クラウド、オンプレミス、エッジを含む計算経路全体を保護することで、現場からデータセンターまで一貫した Confidential AI 環境を構成できる。Confidential AI を一部の先進的な事例から、あらゆる産業が利用できる共通基盤へ発展させるためには、その整備と普及が重要な前提となるだろう。

CHAPTER 06

おわりに

本稿では、「秘密を守れる AI」に求められる信頼性を、環境の信頼性と振る舞いの信頼性の二つに整理した。その上で、Confidential Computing（CC）を中核として AI セキュリティ技術や技術的・運用的統制を組み合わせ、両面から「秘密を守れる AI」を具現化するパラダイムを Confidential AI と定義した。

CC は、TEE による計算中の秘密の保護と、Remote Attestation による実行環境の検証を担う。一方、有害な出力、権限の逸脱、不正利用、学習結果からの情報漏洩には、アラインメント、ガードレール、ハーネス、認証・認可、監視・監査などを組み合わせて対処する必要がある。Confidential AI では、利用形態と脅威モデルに応じて、二つの信頼性を一体的に設計することが重要となる。

CC の進展により、その保護対象は、大規模なフロンティアモデル、複数のエージェント、フィジカル AI へと広がりつつある。今後は、CC 対応 AI データセンターを含む計算基盤と、秘密を開示せずに運用状況を確認する検証可能な透明性を整備し、Confidential AI を幅広い産業で利用できる共通基盤へ発展させることが重要となる。

参考文献

[1] Apple Security Research. Private Cloud Compute Security Guide. https://security.apple.com/documentation/private-cloud-compute/, 2024. Accessed: July 2026.

[2] K. Aubrey. Inside the NVIDIA Vera Rubin Platform: Six New Chips, One AI Supercomputer. NVIDIA Technical Blog, Jan. 2026. Accessed: July 2026.

[3] N. Carlini, N. Cheng, K. Lucas, M. Moore, M. Nasr, V. Prabhushankar, W. Xiao, H. Angulu, E. B. Asher, J. Bow, K. Bradwell, B. Buchanan, D. Forsythe, D. Freeman, A. Gaynor, X. Ge, L. Graham, K. Guru, H. Lakhani, M. McNiece, M. Mehrara, R. Nichol, A. Pirzada, S. Porter, A. Terzis, K. Troy. Assessing Claude Mythos Preview's cybersecurity capabilities. https://www.anthropic.com/research/mythos-preview, 2026

[4] G. Coker, J. Guttman, P. Loscocco, A. Herzog, J. Millen, B. O'Hanlon, J. Ramsdell, A. Segall, J. Sheehy, and B. Sniffen. Principles of remote attestation. International Journal of Information Security, 10(2):63–81, 2011.

[5] Confidential Computing Consortium. A Technical Analysis of Confidential Computing. Technical report, Confidential Computing Consortium, 2023. https://confidentialcomputing.io/wp-content/uploads/sites/10/2023/03/CCC-A-Technical-Analysis-of-Confidential-Computing-v1.3_unlocked.pdf.

[6] Google Cloud. Google Distributed Cloud. https://cloud.google.com/distributed-cloud, 2025. Accessed: July 2026.

[7] H. Liu, C. Shou, X. Liu, H. Wen, Y. Chen, R. Fang, and Y. Feng. Synthesizing Multi-Agent Harnesses for Vulnerability Discovery. arXiv preprint arXiv:2604.20801, 2026.

[8] OpenAI. Patch the Planet: a Daybreak initiative to support open source maintainers. https://openai.com/index/patch-the-planet/. Accessed: July 2026.

[9] M. Sabt, M. Achemlal, and A. Bouabdallah. Trusted Execution Environment: What It is, and What It is Not. In 2015 IEEE Trustcom/BigDataSE/ISPA, volume 1, pages 57–64. IEEE, 2015.

[10] The Linux Foundation. Linux Foundation Launches the Agent2Agent Protocol Project to Enable Secure, Intelligent Communication Between AI Agents. Press release, June 2025.

DIALOGUE

Confidential AI はコンピューターサイエンスの「総合格闘技」— Acompany が描く、R&D の姿と展望

本稿で定義した「Confidential AI」について、開発現場ではどのように議論し、実装しているのか。
Acompany 代表取締役 CEO 高橋亮祐・Chief Scientist 兼 執行役員 VP of Research＆Development 髙橋 翼の二人による対話記事も合わせてご参考いただきたい。

▼対談記事はこちら

サービス紹介はこちら