信頼できるAIシステムの提供に向けたAIリスクマネジメント・AIセキュリティ

AI Risk Management and AI Security for Trustworthy AI Systems

三原　功雄 MIHARA Isao　　小松　みさき KOMATSU Misaki

AIの利活用が拡大する中で，その安全性や信頼性を確保するためのリスク管理が重要な課題となっている。これに対し東芝は，信頼できるAIシステムの実現に向けてAIガバナンス(注1)を構築し，AIリスクマネジメントに取り組んでいる。特に，AIセキュリティに関するリスクへの対応は重要であり，継続的な評価と対策が求められる。しかし，外部提供API（Application Programming Interface）や既製モデルの利用など，内部情報へのアクセスが制限される実務環境では，従来の評価手法だけでは十分でない。

そこで，AIシステムをブラックボックスとして扱い，入出力挙動の分析によりリスク評価を行う“ブラックボックス型AIリスク評価技術”を開発した。この枠組みにより，運用中のAIシステムを含めた継続的なリスク管理を実現する実践的な手法を示した。

In response to the expanding use of artificial intelligence (AI) systems, ensuring safety and reliability of their operation through appropriate risk management has become an issue of critical importance. To realize trustworthy AI systems, Toshiba Corporation has taken the initiative in establishing AI governance and engaging in AI risk management. In particular, continuous evaluation and countermeasures are required to address AI security risks. However, conventional assessment methods are insufficient in practical environments that use external application programming interfaces (APIs) and existing models, where access to internal information is restricted.

To rectify this issue, we have developed a black-box AI risk assessment method that assesses risks through input–output behavior analysis by treating AI systems as black boxes. This method enables continuous risk management of AI systems, including those in operation.

1. まえがき

現在，DX（デジタルトランスフォーメーション）は世界的な潮流となっており，デジタル化の進展に伴い，AI技術の開発及び活用の重要性は一層高まっている。東芝グループでも，DXを通じてカーボンニュートラルやサーキュラーエコノミーの実現に貢献することを目指し，様々な社会課題の解決に向けてAIの適用を進めている。

一方で，AIの利活用が拡大するにつれ，AIの悪用や意図しない動作に起因するトラブルが社会的な問題として顕在化してきた。これに伴い，AIの倫理，ガバナンス，並びに信頼性を確保するための取り組みの必要性が，国内外で強く認識されるようになっている。

国内では，「人工知能関連技術の研究開発及び活用の推進に関する法律」が成立し，内閣府の「人間中心のAI社会原則」を基盤として，総務省及び経済産業省が合同で「AI事業者ガイドライン」を策定するなど，AIに関する原則やルール整備が本格化している。海外でも，欧州AI法（EU Artificial Intelligence Act）をはじめとするAI関連法規制の動きが加速している。これらの動向を踏まえると，AIガバナンスへの対応は，AIを積極的に活用する企業にとって不可欠な経営課題である。

これに対し東芝では，AIガバナンスを構築し，AIリスクマネジメントに取り組んでいる。更に，その重要な一要素であるAIセキュリティに関するリスクについて，AIシステムをブラックボックスとして取り扱い，入出力挙動のみに基づいてリスク評価を行うブラックボックス型AIリスク評価技術を開発した。この技術により，内部情報が十分に得られない場合でも，AIシステムの評価が可能となる。

ここでは，東芝グループにおけるAIガバナンスとAIリスクマネジメントの概略を述べた後，AIセキュリティにおけるリスク対策に必要なブラックボックス型AIリスク評価技術について述べる。

2. 東芝グループのAIガバナンスとAIリスクマネジメント

東芝グループでは，AIを活用した製品・サービスを社会に提供する企業として，AIがもたらす価値を最大化すると同時に，その利用に伴うリスクを適切に管理することが重要であると考えている。この考え方に基づき，AIに対する基本的な

(注1)　AIの利活用によって生じるリスクをステークホルダーにとって受容可能な水準で管理しつつ，そこからもたらされる正のインパクトを最大化することを目的とする，ステークホルダーによる技術的，組織的，及び社会的システムの設計及び運用。

姿勢を示したAIガバナンスステートメント[1]を公表し，AIガバナンスへの取り組みを明確化している。更に，この考え方をベースとして，信頼できるAIシステムの開発・提供・運用を実現するためのAIガバナンスの体制を構築し，具体的な取り組みを推進している。

図1に示すように，東芝グループのAIガバナンスは，AI人材の育成，AI技術の開発・利活用，AIシステムの品質の維持・向上，並びに安全性・信頼性の確保といった複数の要素から構成されており，これらを有機的に連携させることで，AIの価値創出とリスク管理の両立を図っている。

これらの取り組みの中核を成すのが，AIシステムに内在するリスクを把握し，事業判断や開発・運用プロセスに反映させるAIリスクマネジメントである。以下では，まずAIリスクマネジメントの重要性について述べた上で，その具体的なプロセスを示す。

2.1　AIリスクマネジメントの重要性

AIシステムが社会から信頼され，その利活用を持続的に拡大していくためには，AIガバナンスの状態を把握し，継続的な改善を図ることが重要である。個々のAIシステムがもたらす影響とリスクを明確化し，事業判断に資する形で管理する仕組みがAIリスクマネジメントである。

AIシステムの活用では，その有用性や効率向上といった正の側面に注目しがちである一方で，プライバシー侵害や不公平な判断，誤作動による不利益といった負の側面にも十分に目を向ける必要がある。東芝グループにおけるAIリスクマネジメントは，これらのリスクを適切に低減することで，AIシステムの安全性及び信頼性を確保し，社会及び顧客に対して価値の高いAIを提供することを目的としている。

2.2　AIリスクマネジメントのプロセス

東芝グループでは，AIシステムに内在するリスクを体系的に把握し，適切な対応を講じるため，独自のAIリスクアセスメントシートを用いたAIリスクマネジメントのプロセスを構築している。

このプロセスでは，まずAIシステムの開発者が，対象となるAIシステムの基本情報や想定される利用形態について記入し，インシデント発生時の影響度や発生可能性を評価する。これらの回答に基づいてリスクの度合いが整理され，リスク低減に向けた対応方針が提示される。

開発者は，この評価結果を踏まえ，設計の見直しや運用上の対策を検討・実施する。評価と対応を繰り返すことで，AIシステムに内在する潜在的なリスクを段階的に低減していく点に，このプロセスの特長がある（**図2**）。

AIリスクマネジメントでは，品質，プライバシー，公平性，透明性といった多角的な観点からリスクを評価するが，その中でも，悪意ある攻撃や不正利用，情報漏えいといった脅威への対策は，AIシステムの信頼性を確保する上で極めて重要

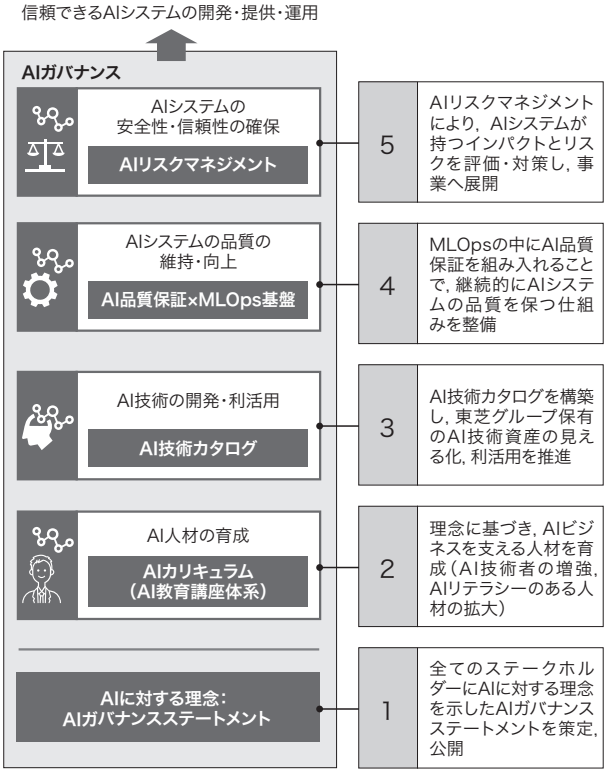

図1．東芝グループのAIガバナンスの全体像

東芝グループでは，AIガバナンスステートメントで示した考え方をベースとして，信頼できるAIシステムの開発・提供・運用に向けて，AIガバナンスを構築し，様々な取り組みを進めている。

Overall architecture of Toshiba Group's AI governance

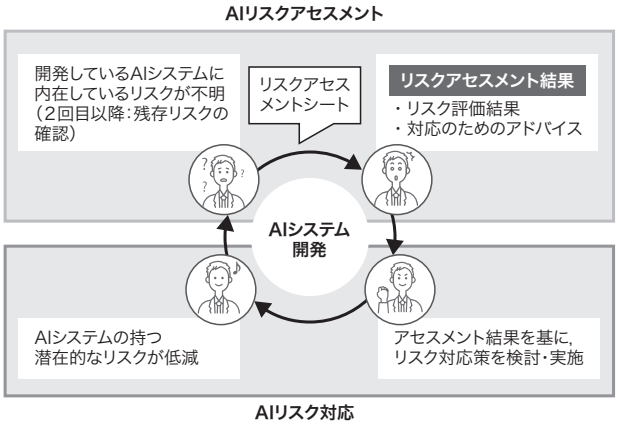

図2．AIリスクマネジメントのプロセス

AIリスクアセスメントとAIリスク対応を繰り返して，AIシステムに内在する潜在的なリスクを段階的に低減していく。

AI risk management process

である。これらの脅威は，従来のIT（情報技術）システムと同様に，あるいはAI特有の特性を踏まえた形で，顕在化する可能性がある。

このため，AIリスクマネジメントを実効性あるものとするには，AIシステムを取り巻くリスクのうち，とりわけセキュリティに関わるリスクを適切に識別し，専門的な対策へとつなげることが不可欠である。3章では，AIリスクマネジメントと密接に連携しながら，AIシステムの安全性を技術的・運用的に確保するためのAIセキュリティの取り組みについて述べる。

3. AIセキュリティの対策に必要なリスク評価技術

ここでは，AIセキュリティに関して，実務環境で継続的なリスク評価を実施するため，どのような評価技術が求められるかを整理する。

3.1 AIセキュリティとは

AIセキュリティは，学習データの汚染，モデルの抽出・改ざん，入力を通じた出力操作，出力を介した情報漏えいなど，AI技術固有の構成要素や振る舞いに起因する脅威を扱うセキュリティ分野である。これらの脅威は，AIシステムの入出力や学習過程と強く結びつくため，既存のセキュリティ技術を前提としながらも，AIシステム特有の攻撃面と影響を踏まえた評価が必要となる。

そのため，AIセキュリティでは，システムの構成や運用状況を踏まえ，どのような攻撃が成立し得るか，またその攻撃がどのような影響を及ぼし得るかを継続的に評価することが重要となる。3.2節以降では，このようなリスク評価に関する現行の取り組みと，そこに残る課題を整理する。

3.2 現行のリスク評価への取り組み

AIシステムのリスク評価に向けた取り組みの多くは，評価者が評価対象システムの設計情報や学習データなどの内部情報へ継続的にアクセスできること（ホワイトボックス）を暗黙の前提としている。そのため，内部情報が十分に得られない場合，評価自体が難しくなる。

また，AIシステムの脆弱（ぜいじゃく）性を外部から検知する方法として，攻撃者の視点から評価するレッドチーミングも注目されている。しかし，その実施には高度な専門知識や，対象システムの振る舞いに関する十分な理解が必要であり，コストや実施負担が大きいという問題点がある。

3.3 実務環境で生じ得るギャップ

近年，外部提供APIや既製モデル，分業開発などを前提とした開発形態が増えている。こうした状況では，AIシステム提供者が，リスク評価に必要な内部情報を把握できるとは限らない。

更に，AIシステムは運用を続ける中で入力データの傾向や利用環境が変化する。そのため，状況に応じて内部のAIモデルの更新や再学習が行われる。その結果，リスクの内容や程度が変化するため，リスク評価は一度きりではなく，必要なタイミングで繰り返し実施できることが必要である。

3.4 ブラックボックス型AIリスク評価の必要性

AIのリスク評価は，モデルの設計情報や学習データを確認できることが望ましい。しかし，3.3節で述べたとおり，内部情報へ恒常的にアクセスできない形態（外部提供API利用や，既製モデル利用，分業開発など）が増えている。加えて，運用中の変更などにより評価を繰り返し行うことが必要となる場合があり，これらへの対応が課題である。

このような状況下で継続的なリスク評価を実現するには，AIシステムをブラックボックスとして取り扱い，入出力挙動だけを用いてリスク評価を行うアプローチが有効と考えられる。

4. ブラックボックス型AIリスク評価技術

ブラックボックス型AIリスク評価技術は，AIシステムをブラックボックスとして取り扱い，入出力の分析によってリスク評価を行う。内部情報（AIモデルや学習データなど）へのアクセスが制限される状況でも実行でき，運用中の任意のタイミングで評価を行える可能性がある。

図3は，ブラックボックス型AIリスク評価技術の基本的な考え方を示している。システム利用者は，AIシステムに情報を入力することで出力を得る。評価者がこの入出力を分析す

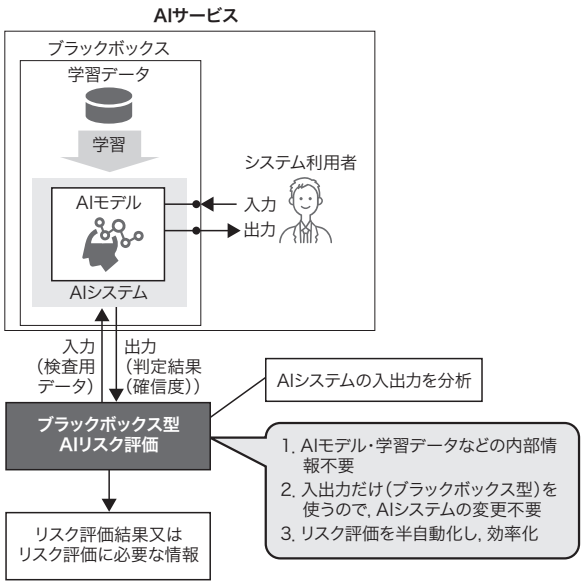

図3. ブラックボックス型AIリスク評価技術の概略

AIサービスをブラックボックスとして，入出力を分析し，内部情報不要でリスク評価結果又はリスク評価に必要な情報を得る。

Overview of black-box AI risk assessment method

ることで，リスク評価結果や評価に必要な情報を得る。4章では，この考え方をデータバイアス検査に適用する例を示し，入出力分析によってどのようにリスクの兆候を把握するかを説明する。

4.1　入出力分析によるデータバイアス検査の例

ブラックボックス型AIリスク評価技術の開発に向け，画像を扱う推論AIを対象に，システム利用者を含む外部からの問い合わせに対する入出力を分析し，AIシステムや学習データの性質に関する情報がどの程度推定され得るかを調査している。具体的には，顔識別モデルの学習データに男女などの属性ごとのデータ数に偏りがある場合，当該顔識別モデルを搭載するAIシステムは，学習データが少ない属性に対して認識精度が低下したり，誤識別率が高くなったりする可能性がある。結果，属性によってサービス利用のしやすさに差が生じるリスク（データバイアスのリスク）がある。ブラックボックス型AIリスク評価技術では，このようなリスクを入出力分析から検出できるように設計する。技術的なポイントは，学習データの偏りそのものを直接確認するのではなく，その偏りがAIシステムの出力挙動にどのように現れるかに着目する点である。具体的には，属性ごとに入力データを準備し，入力に対する出力の変化，属性間の出力差，誤識別傾向などを比較することで，データバイアスを推定する。この考え方は，AIシステムの入出力だけから学習データの属性分布を推定する外部情報漏えい攻撃において用いられる手法・考え方を，攻撃目的ではなく，防御側の検査技術へ応用するものである。

図4は，ブラックボックス型AIリスク評価技術の入出力だけを用いた分析をデータバイアス検査に活用する例である[(2)]。銀行の口座開設などで使用されるオンライン本人確認を例に，学習データの偏りが不公平な出力につながるリスクを，サービスのユーザーAPIから点検することを想定している。

4.2　適用範囲の拡張方針

ブラックボックス型AIリスク評価技術の評価対象は，顔識別モデルに限らず，物体検知モデルなど，別のモデルへの拡張も視野に入れ，検討している。また，評価項目（脅威など）についても，段階的に拡張する方針である。これにより，ブラックボックス型AIリスク評価技術を，特定のモデルに限定されない，多様なAIモデル・サービスに適用可能なブラックボックステスト形式のリスク評価支援ツールとして発展させ，その実用性を高めることを目指す。

4.3　課題

ブラックボックス型AIリスク評価技術によるリスク評価は，AIサービスの提供前だけでなく，提供開始後の運用時でも，環境変化やモデル更新に応じて繰り返し実施できることが重要である。

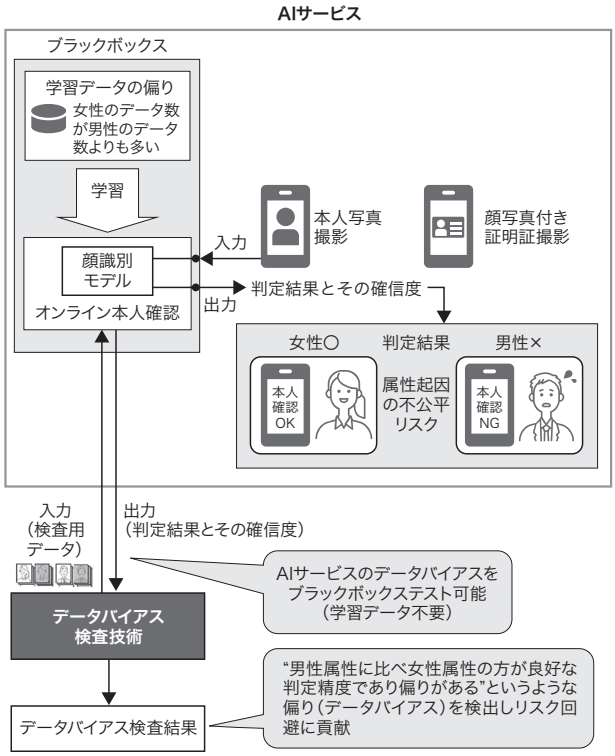

図4. ブラックボックス型AIリスク評価技術による分析をデータバイアス検査に活用する例

オンライン本人確認サービスの顔識別で生じる判定差（属性起因の不公平）を，内部の顔識別モデルの情報を使用せず，ブラックボックス型AIリスク評価技術で検知する。

Example of application of black-box AI risk assessment method to data bias detection

このため，適用範囲の拡張に加え，ブラックボックス型AIリスク評価技術によって得られる結果を，既存のAIリスクマネジメントのプロセス中で，どのような判断材料として扱うかを整理する必要がある。

具体的には，対策要否の判断，対応の選択，再評価の実施といった場面で，ブラックボックス型AIリスク評価技術の結果をどのように活用するかを明確にすることが，課題として挙げられる。

5. あとがき

ここでは，信頼できるAIシステムの開発・提供・運用を実現するための基盤として，AIガバナンスの考え方を整理するとともに，AIリスクマネジメント及びAIセキュリティの位置付けとその具体的な取り組みについて述べた。AIの利活用が拡大し，その形態が多様化する中で，AIに内在するリスクを適切に把握し，継続的に管理していくことの重要性は，今後更に高まっていくと考えられる。

特に，外部提供APIや既製モデルの利用など，内部情報へ

のアクセスが制限される実務環境では，従来のホワイトボックスを前提とした評価手法だけでは十分な対応が困難な場面も想定される。ここで述べたブラックボックス型AIリスク評価のアプローチは，こうした制約下でも，運用中のAIシステムを含めた継続的なリスク把握を可能にするアプローチとして，有力な選択肢となり得る。

東芝グループでは，AIガバナンスの枠組みの下，AIリスクマネジメント及びAIセキュリティの取り組みを相互に連携させながら，社会や顧客から信頼されるAIシステムの実現を目指している。ここで示した考え方や取り組みが，AIを活用する企業・組織における実践的な議論の一助となれば幸いである。

文　献

⑴　東芝. “東芝グループAIガバナンスステートメント”. 東芝のAI. <https://www.global.toshiba/jp/technology/corporate/ai/statement.html>, (参照 2026-04-21).

⑵　東芝. サイバーセキュリティ報告書2026. 2026, 54p. <https://www.global.toshiba/content/dam/toshiba/jp/cybersecurity/corporate/report/pdf/CyberSecurityReport2026_A4.pdf>, (参照 2026-06-18).

三原　功雄　MIHARA Isao
総合研究所 AIデジタルR&Dセンター AI戦略部
ACM・人工知能学会・情報処理学会・電子情報通信学会会員
AI Strategy Dept.

小松　みさき　KOMATSU Misaki
総合研究所 AIデジタルR&Dセンター セキュリティ基盤研究部
Security Research Dept.