Business Software Alliance

コンテンツ真正性の基礎解説

AI生成コンテンツの透明性は、責任あるAIの推進における重要な要素です。

Business Software Alliance（ビジネス・ソフトウェア・アライアンス、BSA）は、ユーザーがAI生成コンテンツの真正性と来歴を特定しやすくすることができる、信頼性の高いコンテンツ認証および来歴確認メカニズムの開発と導入を支持しています。これにより、消費者はコンテンツが人間によって生成されたものか、AIによって生成されたものかを判断しやすくなり、誤情報や偽情報に対処する上でも役立ちます。

www.bsa.org

世界中の政策立案者は、消費者がオンラインで目にする写真や動画が本物かどうかをどう判断できるかという重要な問題に直面しています。そのためには、電子透かし、デジタルフィンガープリント、安全なメタデータなど、今日すでに存在するツールについて、消費者が知る必要があります。これらのツールによって、画像または動画の作成者、また、AIなどのデジタルツールによって変更されたかどうかを明らかにすることができます。政策立案者は、コンテンツが人間によって作成されたものか、AIによって作成されたものかを消費者にわかりやすくするために、こうしたツールの使用を支持すべきです。

コンテンツ真正性に関するBSAの基礎解説では：

» AI透明性政策の基礎となるコンテンツ真正性およびコンテンツ来歴を定義します。

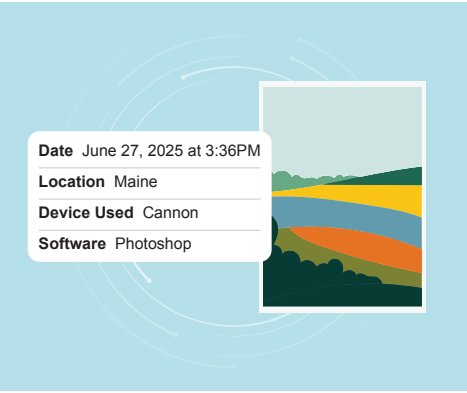
» 画像または動画の作成者を特定し、その画像や動画が変更されているかどうかを識別できる既存のツールについて説明します。これには、機械可読な電子透かし、デジタルフィンガープリント、安全なメタデータなどがあります。

» コンテンツの真正性や来歴確認ツールの利用を促進するために、さまざまなタイプの企業が実施できる各種対策について説明します。

» ディープフェイクによって生じる政策上の課題について考察します。

» オンラインで事実と虚構を見分ける上で、さまざまなAI政策が消費者にどう役立つのかを説明します。

コンテンツ来歴およびコンテンツ認証とは？

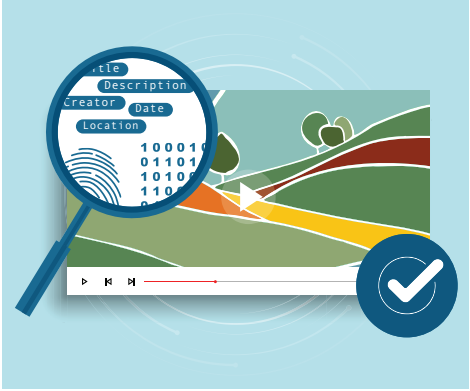


コンテンツ来歴

コンテンツ来歴情報は、コンテンツの出所と変更履歴を示すもので、画像、動画、音声クリップなどのデジタルファイルの作成元、所有権、および変更履歴を追跡します。例：写真には、写真が撮影された日付と場所、撮影に使用されたカメラの種類、およびファイルの編集に使用されたソフトウェア編集プログラムを示す、暗号署名されたメタデータが含まれています。

コンテンツ認証

コンテンツ認証は、コンテンツとその来歴情報が信頼できるかどうか、およびコンテンツのメタデータ、透かし、またはクレデンシャル（認証情報）が改ざんされていないかどうかを示すのに役立ちます。例：コンテンツ認証ツールは、動画の埋め込み署名をチェックして、発行者の公開鍵と一致することと、ファイルが公開後に変更されていないことを確認できます。

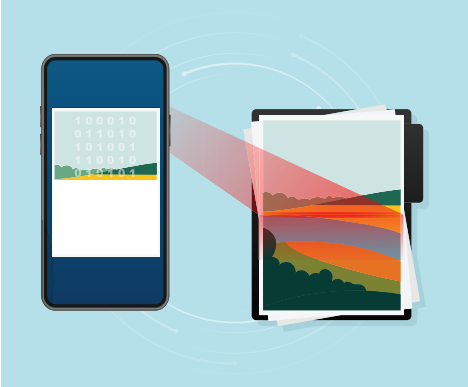


AI生成コンテンツの識別に役立つツール

AIツールによって生成されたコンテンツを含め、コンテンツがどのように作成または変更されたかを特定できる技術には、さまざまなものがあります。政策立案者は、単一の手法の使用を義務付けるのではなく、企業がさまざまな技術を使用して、コンテンツがAIによって生成されたものかどうかを示すことができるようにすべきです。

機械可読な電子透かし

機械可読な電子透かしは、専用ツールによって検出可能な少量の情報を埋め込むことで、ファイルに直接追加されます。電子透かしは、ファイルのピクセルまたはデータストリームに埋め込むことができるため、後でメタデータが削除された場合でも、ファイルの出所を確認できます。例：画像内の不可視透かしは、その画像を作成者または真正性プラットフォームに結びつけることができます。

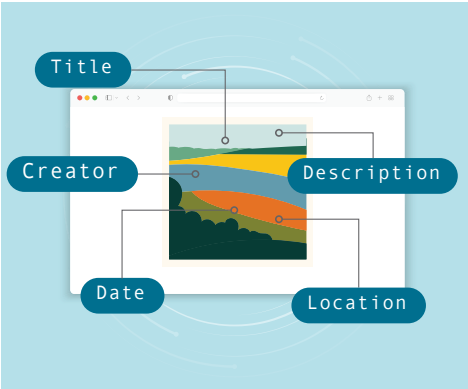


デジタルフィンガープリント

デジタルフィンガープリントは、色のパターン、構造、エンコードなど、コンテンツの核心的な特徴から固有の識別子（つまり「フィンガープリント」）を生成するものです。デジタルフィンガープリントは、ファイルを変更するのではなく、ハッシュなど、ファイルとは別に格納できる固有の数学的署名です。ファイルのデジタルフィンガープリントをデジタルフィンガープリントのデータベースと照合することで、ファイルがオリジナルと一致するかどうか、またはファイルが変更されているかどうかを確認できます。例：動画をフレームに分割して、視覚的および音声的手がかりからフィンガープリントを作成できます。その署名をデジタル署名のデータベースと照合することで、動画が変更または改変されたかどうかを確認できます。

安全なメタデータ

安全なメタデータを使用すると、ファイルの作成者、作成に使用されたツール、およびファイルにどのような編集が加えられたかに関する情報を格納できます。この情報により、ファイルのコンテンツ来歴が証明されます。例：画像には、作成者の身元、編集履歴、使用ソフトウェアを記録するメタデータを含めることができます。これは、閲覧者がその真正性と出所を確認できるように、デジタル署名されてファイルに埋め込まれます。



堅牢なアプローチ：複数のツールの組み合わせ

これらの技術を連携させることで、コンテンツ真正性に対する堅牢なアプローチを支え、コンテンツの作成方法や改変の有無を確認するためのツールを消費者に提供することができます。3 つのプロセスはすべて検証レイヤーとして機能します。

電子透かし
真正性を埋め込み

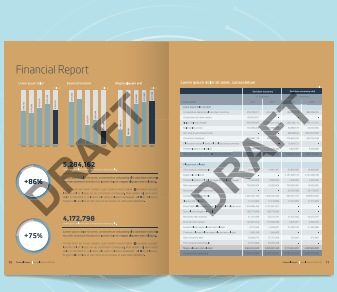
フィンガープリント
コンテンツを識別

メタデータ
コンテキストと作成者を記録することが可能

多くの状況で 3 つのプロセスをすべて実装することが理想的ですが、常にすべての技術をすべての場面で使用する必要はありません。むしろ政策上は、今すぐ採用可能なほど十分に成熟した技術的手法の使用を奨励しつつ、多様な技術的・運用上の用途に対応するのに十分な柔軟性を維持すべきです。

AI との相性の悪さ：可視透かし

可視透かしは、文書が「ドラフト」または「機密」であることを明確に示すなど、コンテンツのラベル付けに長年使用されてきたローテクな方法です。AI が普及した現在、可視透かしは簡単に削除可能であることから、これを義務付けることは実際的ではなく、むしろその有効性を損ない、誤った安心感を生み出します。



注目

AI 生成テキスト

画像や視聴覚コンテンツにラベルを付けて、そのコンテンツが本物か AI によって生成されたものか消費者が判断しやすくするツールには、さまざまなものが存在します。その一方で、AI 生成テキストへのラベル付けを義務付けることは、多くの実務上の懸念を引き起こします。消費者がテキストベースの AI 生成コンテンツを確実に見分けることができるようにするために、やりとりをしている相手が AI システムであることを消費者が認識できるようにする要件に重点を置くことを提言します。

注目

C2PA (Coalition for Content Provenance and Authenticity／コンテンツ来歴および信頼性のための標準化団体) 規格

C2PA規格は、製品やプロセスにデジタル来歴情報を組み込むためのもので、誰でも使用できます。いくつかの技術を組み合わせており、ISO (国際標準化機構) によって国際規格として承認される見込みです。

C2PA規格の仕組み

1. **ファイルの作成**
ユーザーは、新しいファイルを作成する際、ファイルの作成方法や編集内容など、ファイルの来歴に関する情報を生成するツールを使用できます。

2. **コンテンツクレデンシャル**
来歴情報をコンテンツクレデンシャルに埋め込みます。

3. **暗号署名**
コンテンツクレデンシャルの真正性を保証するために、その作成に使用したソフトウェアまたはハードウェアの秘密鍵で、コンテンツクレデンシャルは署名されます。公開鍵は認証のために提供されます。

4. **埋め込みおよび／または透かし埋め込み**
コンテンツクレデンシャルは、ファイル内への埋め込みや、不可視透かしまたはデジタル署名に関連付けられている別のデータベースへの格納など、複数の場所に格納できます。

5. **検証**
ファイルの真正性を確認するには、コンテンツクレデンシャルを確認するツールを使用して、ファイルが本物か変更されているかどうかを確認できます。

6. **表示**
真正性が検証されたコンテンツは、バッジやアイコンなどの明確な表示を付けることができます。

コンテンツクレデンシャルとは？

最新のC2PA規格では、複数の手法を組み合わせたコンテンツクレデンシャルを利用しています。

» コンテンツクレデンシャルには、ファイルの作成日時や作成に使用された技術など、ファイルに関するメタデータが含まれています。

» すでにメタデータが含まれているコンテンツクレデンシャルに、透かし識別子やデジタル署名を追加することができます。

» クレデンシャル (認証情報) のパッケージはデジタル署名され、そのファイルに一意に関連付けられます。

» コンテンツクレデンシャルは、コンテンツに埋め込まれるとともに、コンテンツクレデンシャルのデータベースなどに別途格納されます。

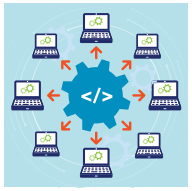
AIサプライチェーン:さまざまな役割と責任

AIサプライチェーンは常に進化しており、さまざまな企業が関わっています。立法者は、コンテンツ真正性に関する政策を立案する際、これらのさまざまなタイプの企業の存在を認識していなければなりません。なぜなら、それらの企業はそれぞれアクセスできる情報も異なれば、消費者保護対策に対する立場も異なるからです。

例えば、ある企業がAIモデルを開発し、別の企業がそのAIモデルをアプリケーションに統合し、さらに別の企業がそのアプリケーションを使用して新しいコンテンツを作成することがあります。

モデル開発者

モデル開発者は、さまざまなアプリケーションで使用できるAIモデルを開発します。例えば、ある企業がさまざまなユースケースに適応できる基盤モデルを構築すると、その同じ基盤モデルを使用して、検索エンジン、チャットボット、スパム検出ソフトウェア、長文テキストの要約ツールなどのさまざまなアプリケーションを強化することができます。モデル開発者は、その基盤モデルがどのように開発されたかについての情報は持っていますが、他の企業がそのモデルを特定の用途にどのように展開しているかについての情報は通常持ち合わせていません。



インテグレーター

インテグレーターは、特定のアプリケーションにAIモデルを統合して、他の企業が使用できるようにします。インテグレーターによっては、AIモデルを特定のサービスやアプリケーションに接続するだけの場合もあれば、AIモデルをファインチューニングまたは変更した上でサービスやアプリケーションに搭載する場合もあります。インテグレーターは通常、AIモデルに加えた変更についての情報は持っていますが、モデルの初期開発や、結果として得られたAIアプリケーションを他の企業が使用する特定の状況に関する直接の知見は持ち合わせていません。例えば、1つ以上のAIモデルを組み込んだAIアプリケーションを開発する企業は、インテグレーターとしての役割を果たすことがあります。

導入者

特定の目的にAIツールを使用する企業は、多くの場合、導入者と呼ばれます。導入者は、特定のAI技術をいつ、どのように使用するかを決定するため、特定のユースケースの事実について知見を得ることになります。しかし、他の企業からAIツールを入手することが多いため、通常はAIツールの初期トレーニングに関する直接の知見を持ち合わせていません。

コンテンツ真正性に関する政策には、これらのさまざまな役割を反映する必要があります。

これらの大きく異なるタイプの企業に画一的な要件を適用する政策は、消費者を効果的に支援することにはなりません。例えば、コンテンツを生成するAIアプリケーションの開発者は、そのAIアプリケーションによって生成されるコンテンツにコンテンツ来歴情報を加えるのに最適な立場にあります。対照的に、基盤モデルの開発者は、他の主体によって開発および使用されるAIアプリケーションによって生成されるコンテンツに、電子透かしなどのデジタルマーキングを加える技術能力を欠いていることになります。

時代に合わせた法律の適切性の確保: コンテンツ来歴を保証するソリューションにおける最先端技術を構成するものは、時間と共に進化します。政策立案者は、あらゆる法的枠組みがそうした進展に対応できるようにすべきです。C2PAのようなオープン規格の採用は、この取り組みに役立ちます。

ディープフェイクとの戦い

不可視透かし、デジタルフィンガープリント、安全なメタデータなどのツールは、真正なコンテンツを識別するのに役立ちます。それにより、消費者は自分が接する画像、動画、音声クリップが本物かどうかを判断することができます。

これらが相まって、ディープフェイクの問題に対処するのに役立ちます。悪意のある行為者が自分のディープフェイクに「フェイク」と表示するとは思えませんが、重要なのはコンテンツの真正性を判断できるツールを消費者に供することです。そのためには、これらのツールがさまざまなデバイスやプラットフォームで採用される必要があります。また、消費者がオンラインで接する情報に対して健全な懐疑心を持ちながら、画像や動画の来歴に関する情報に注意する必要があることを理解できるように、これらのツールについて啓発することも必要です。

悪意のある行為者は、今後も人を欺く目的でAIなどの技術を悪用する新たな方法を見つけ続けるでしょう。しかし、安全なメタデータやC2PA規格のようなツールは、善良な行為者がコンテンツの真正性を証明する上で不可欠であり、全国民がオンラインで事実と虚構を見分けるのに役立ちます。

政策立案者は、ディープフェイクに対処する最も効果的な方法を検討する際に、さまざまなタイプの企業が持つさまざまな役割と機能も考慮に入れるべきです。これは、サービスレベル、技術、機能、ユーザーベースの主な違いにより、さまざまなタイプの企業が存在し、それぞれ異なる問題に対処する立場にあるため、非常に重要です。

また、政策立案者はさまざまなタイプのサービスに付随するさまざまなリスクも認識すべきです。例えば、B2B (business-to-business) ソフトウェアサービスは、ユーザーベースの規模や消費者に直接サービスを提供しないという事実を踏まえると、ユーザーの安全や公の秩序に対するリスクは限定的であり、ディープフェイク関連の懸念はあまり生まないかもしれません。

AI政策による消費者支援のあり方

消費者がオンラインで事実と虚構を見分けることができるようにしたいと考える政策立案者は、コンテンツがAIによって生成されたものかどうかを識別できる既存のツールを利用することが重要です。

政策目標	ソリューション
 消費者の対話相手がAIシステムであることを知らせる。	企業が消費者と直接対話するAIシステムを提供する場合、一目瞭然でない限り、対話相手がAIシステムであることを消費者に伝えるべきである。
 消費者がAI生成コンテンツを見分けやすくする。	コンテンツの来歴確認ツールや認証ツールの使用義務は、消費者向け音声、画像、または動画コンテンツに重点を置くべきである。これらの要件は、テキストまたはB2Bの状況には適用すべきではない。これが重要なのは、消費者がAIツールを使用して電子メールのテキストを編集したり、文書を翻訳したりする場合のように、テキストにはコンテンツ真正性メカニズムを使用する意味がないからである。B2Bの状況では、企業は特定の用途に基づいてコンテンツ真正性に対処する方法を別途策定することができる。
 AI生成コンテンツを識別するための世界の先進技術の使用を支援する。	AI生成コンテンツの識別要件は、C2PA（Coalition for Content Provenance and Authenticity／コンテンツ来歴および信頼性のための標準化団体）などのオープン規格をはじめとする世界の主要なツールによって確実に満たすことができる。多くの国がコンテンツ来歴に関する規制を検討しているが、各国固有の規格を採用すると、随時更新されながら広く受け入れられている規格を消費者が利用できなくなる可能性がある。
 AI生成コンテンツに関する透明性を促進する。	AIアプリケーションはコンテンツの生成に使用されるため、AI生成コンテンツにコンテンツ来歴情報を加える義務は通常、AIアプリケーションの開発者に課すのが最も適切である。企業は、機械可読な電子透かし、デジタルフィンガープリント、安全なメタデータなどのツールを使用してコンテンツの出所を特定することができる。
 ユーザーが常にコンテンツ来歴情報を利用できる状態を確保する。	セキュリティ上の懸念がない限り、機械可読な電子透かしや安全なメタデータなどのコンテンツ来歴情報の削除を禁止する。