令和7年度

インターネット上の偽(にせ)・誤(ご)情報等への対策技術の開発・実証事業 成果発信イベント

～偽・誤情報のリスクへ立ち向かう技術の最前線～

日時 2026年3月16日(月) 13:00～17:00

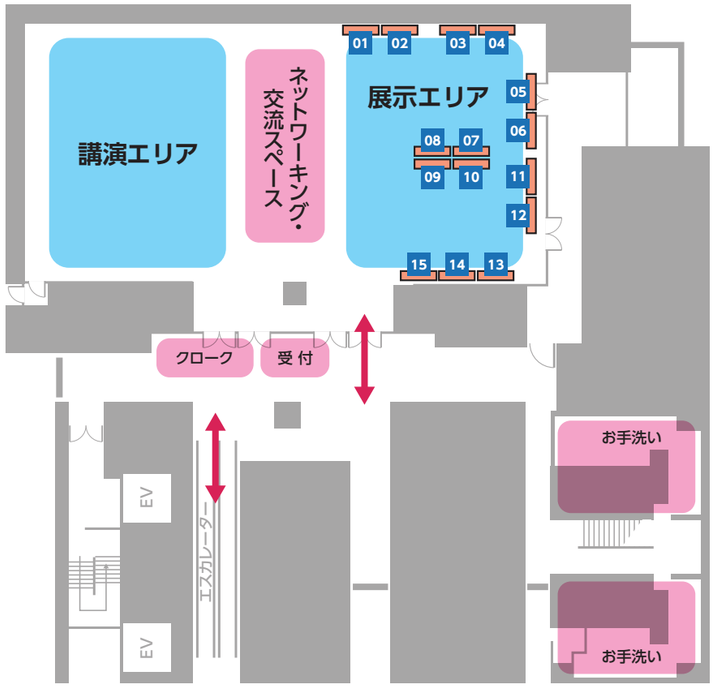
会場 東京・大手町サンケイプラザ

主催 総務省

総務省事業
令和7年度
インターネット上の偽・誤情報等への対策技術の開発・実証事業 成果発信イベント
～偽・誤情報のリスクへ立ち向かう技術の最前線～

主催：総務省
HP　：https://www.soumu.go.jp/

1 会場MAP・出展者一覧



本イベント会場内における
撮影、録画・録音はご遠慮ください。

■ ブース出展者一覧

ブースNo	社　名
01	一般社団法人コード・フォー・ジャパン
02	エヴィクサー株式会社
03	NTTドコモビジネス株式会社
04	Originator Profile技術研究組合
05	株式会社Classroom Adventure
06	株式会社コンステラセキュリティジャパン
07	株式会社TDAI Lab
08	株式会社データグリッド
09	関西テレビソフトウェア株式会社
10	SEARCHLIGHT株式会社
11	Sakana AI株式会社
12	サン電子株式会社
13	NABLAS株式会社およびNTT東日本株式会社
14	日本電気株式会社
15	偽・誤情報対策に係る研究・調査主体 (株式会社新領域安全保障研究所、 東京大学大学院情報学環)

2 講演スケジュール

TIME TABLE	CONTEANT		
12:30～13:00	客入れ		
13:00～13:05	開会挨拶		
13:05～13:30	基調講演	インターネット上の偽・誤情報対策技術の現在地とこれから 笹原 和俊 教授（東京科学大学）	
13:30～13:45	休　憩		
13:45～13:50	ショートプレゼン 前半	01	一般社団法人コード・フォー・ジャパン
13:50～13:55		02	エヴィクサー株式会社
13:55～14:00		03	NTTドコモビジネス株式会社
14:00～14:05		04	Originator Profile技術研究組合
14:05～14:10		05	株式会社Classroom Adventure
14:10～14:15		06	株式会社コンステラセキュリティジャパン
14:15～14:20		07	株式会社TDAILab
14:20～14:30			休　憩
14:30～14:35	ショートプレゼン 後半	08	株式会社データグリッド
14:35～14:40		09	関西テレビソフトウェア株式会社
14:40～14:45		10	SEARCHLIGHT株式会社
14:45～14:50		11	SakanaAI株式会社
14:50～14:55		12	サン電子株式会社
14:55～15:00		13	NABLAS株式会社およびNTT東日本株式会社
15:00～15:05		14	日本電気株式会社
15:05			

3 本開発・実証の概要

インターネット上の偽（にせ）・誤（ご）情報等への対策技術の開発・実証事業
（令和7年度）

総務省は、生成AIに起因する偽・誤情報を始めとした、インターネット上の偽・誤情報の流通・拡散に対応するため、「インターネット上の偽・誤情報等への対策技術の開発・実証事業」を通じ、対策技術の開発・実証及び社会実装を推進することとしています。

本開発・実証では、公募により採択された、以下の技術開発主体14者、研究・調査主体6者により事業が実施されました。

技術開発主体	研究・調査主体
一般社団法人コード・フォー・ジャパン	株式会社新領域安全保障研究所
エヴィクサー株式会社	中央大学
NTTドコモビジネス株式会社	東京大学大学院情報学環
Originator Profile技術研究組合	東京大学大学院工学系研究科
株式会社Classroom Adventure	名古屋工業大学
株式会社コンステラセキュリティジャパン	日本エンタープライズ株式会社
株式会社TDAI Lab	
株式会社データグリッド	
関西テレビソフトウェア株式会社	
SEARCHLIGHT株式会社	
Sakana AI株式会社	
サン電子株式会社	
NABLAS株式会社およびNTT東日本株式会社	
日本電気株式会社	

本事業の技術区分のご紹介

偽・誤情報対策技術とは?

技術区分 Ⅰ

コンテンツの真偽判別支援・改ざん検知技術

情報の受信者が、
インターネット上の情報が本物であるか、
改ざんされていないか
見極めることの支援

技術区分 Ⅱ

真正性保証・信頼性判断支援技術

情報コンテンツの作成者・発信者が
本物であることを示し、
情報コンテンツの信頼性の
判断を支援

技術区分 Ⅲ

情報流通情報の可視化・分析技術

インターネット上の
情報の広がりを把握する技術

技術区分 Ⅳ

情報の拡散防止・無効化技術

偽・誤情報の拡散を未然に防ぎ、
影響を抑える技術

一般社団法人コード・フォー・ジャパン

SNSにおける偽情報・真偽不明情報の市民参加型可視化・分析技術

1 背景

日本では、SNS上の偽・誤情報に関する定量的研究やツール開発が海外に比べて少ない。
その主な原因として、2023年のX API有料化が挙げられる。日本は特にX利用者が多く、学術研究向けAPIに依存していたため、その影響は大きい。

2 目的

本事業で開発中のBirdXplorerは、Xのコミュニティノートのデータを活用し、可視化やデータ提供を行うことで、SNS上の定量的な研究・報道を促進する。
また、災害や選挙時に偽情報・真偽不明情報を迅速に提供し、市民が主体的にSNS検証に取り組む環境の構築を目指す。

3 開発技術の概要

本事業で開発する「BirdXplorer」は、X（旧Twitter）のコミュニティノートのデータを活用した偽情報・真偽不明情報の可視化ツール。
コミュニティノートは、一般ユーザーが投稿に対して情報を補足する機能で、偽情報や真偽不明情報に対して使われることが多い。コミュニティノートのデータをもとにグラフ等のビジュアライゼーションが掲載されたダッシュボードを提供する。ダッシュボードではコミュニティノートの時系列的な変化やナラティブを可視化する。

4 社会実装のイメージ

2025年度は報道関係者や研究者に限定して提供。2026年度の一般公開を目指して開発を進める。

- **想定しているユーザ** → 報道機関、研究者、一般市民
- **具体的なユースケース** → コミュニティノートの傾向把握、ナラティブの傾向把握
- **ユーザ・導入先への提供形態** → 無料での提供
- **ビジネスモデル** → 法人や個人による寄付による収入

Code for Japan

エヴィクサー株式会社

音響透かしと音響フィンガープリントを用いた偽・誤情報対策クラウドシステムの開発・実証

1 開発・実証の背景・目的

背景
生成AIによる偽音声、ディープフェイクの急増
↓
音声領域の真偽対策が急務

目的
音響透かし・AI判定等を統合したシステムを開発
↓
「音声の真正性」を保証

2 開発技術の概要

三つの技術を統合した
「Evixar Audio Forensics Blocks」

1. **音響透かし**
   来歴情報の埋込（真正性保証）
2. **音響フィンガープリント**
   特徴量照合（改ざん検知）
3. **AI合成音声判定**
   生身の音声とAI合成音声の判定

3 今年度の開発アプローチ

社会実装のための機能強化

1. **UI/UXの改善と判定精度向上**
   最新生成AIモデルへの対応強化
2. **モジュール化**
   ニーズに即した機能提供に対応
3. **インセンティブ設計のPoC**
   LINE連携によるキャンペーン機能の実装
4. **社会インフラへの対応**
   電話等でのリアルタイム処理実現

4 実証実験の成果

1. **AI合成音声判定**（データアナリティクスラボ株式会社と共同研究）
   様々な合成モデルに対し、高い判定精度を達成
2. **大規模フィールド実証**
   Bリーグ等にて数千人規模で機能提供
3. **リアルタイム処理性能の向上**
   IP電話やIPサイマルラジオサービスでの実装に成功
4. **東京都が表彰、新事業分野開拓者に認定**
   政策目的随意契約が可能に

5 進行中の社会実装

行　政	首長会見等の自治体の公式動画に真正性保証
企業向け トラストサービス	シヤチハタ株式会社と提携し「音のしるし」サービス（ODM）を展開、社会実装を加速
通　信	詐欺電話をインフラ側で遮断
放送・メディア	防災・災害情報の来歴保証、不正流通トラッキング

「音」の独自技術で
安心出来る社会をつくる
新しい情報インフラの構築

SIGNED SOUND

evixar

NTTドコモビジネス株式会社

情報の真正性を可視化するC2PA技術を活用した偽・誤情報対策の開発・実証

1 開発・実証の背景・目的

生成AI等の普及により現実との区別が難しい加工コンテンツが増加、第三者から取得したコンテンツを業務利用する際のファクトチェックの稼働が増加している。本開発技術においてファクトチェック業務の削減を目指す。

2 開発技術の概要

スマートフォンのカメラを使用し画像・動画を撮影時に、撮影デバイスの真正性チェック機能とC2PA方式での正しい撮影時のメタデータ付与機能を提供。また、本技術によりメタデータの事前チェックを行いその結果を付与する。

3 開発技術の有効性（協力事業者との実証実験）

開発した真偽判定支援ツールの有効性について、ファクトチェックを行っている事業者2社から協力を得て、疑似のSNS投稿を準備し内容が真実であるか、データの加工がされているか確認する実証実験を実施。

実証実験の結果、情報の真偽判定結果における**検知率としては85%以上**、ファクトチェックに要する時間としては**15%以上の稼働削減**を達成した。

協力事業者とのヒアリングから、情報の取り扱いの拡充及び情報の信頼性確保に繋がるという意見を確認した。

4 社会実装のイメージ

報道業界を始めとする画像・動画のファクトチェックを実施または必要とする業界に対して、画像・動画の真偽判定支援ツールとしてクラウドサービス化し、コンテンツの信頼性およびC2PA署名されている情報を確認できるサービスを提供することを想定している。

また、国内外の端末メーカーやOSベンダーとの協議を深化させ、スマートフォンカメラへの真正性付与機能の標準搭載（デフォルト化）を目指す。

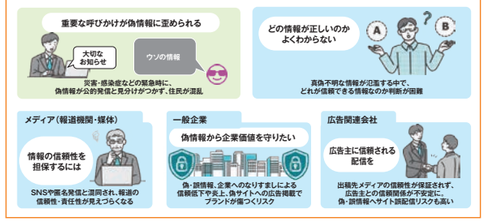
docomo business

Originator Profile技術研究組合

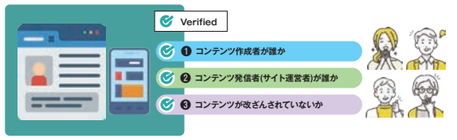
「Originator Profile」の開発と社会実装

1 開発・実証の背景・目的（解決したい社会課題、目指す姿）

開発技術によりアプローチする課題（インターネット空間の様々な課題）



課題を踏まえ目指す姿・ゴール（エンドユーザーがOPによって確認できること）

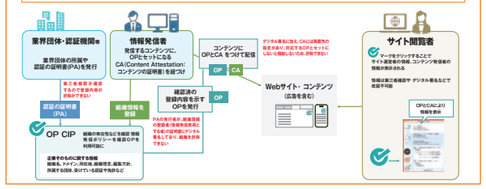


コンテンツ作成者・発信者が加盟する業界団体などの第三者機関が、組織の存在などを確認し、その情報をユーザーがブラウザで検証できるようになります。
ユーザーにとっては、コンテンツの作成者・発信者が確認できることにより、偽・誤情報が判別しやすくなります。

2 開発技術の概要（技術開発の取組・成果）

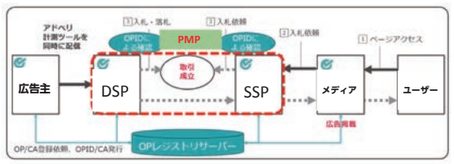
実証実験で開発したOPシステム概要

仮のPAを含む検証用OPSを発行し、CAサーバーも立てて、メディアのCMSが適切にOPSとCASをコンテンツに実装できるように実証を行った。



開発した広告関連システムと実証実験

広告主、DSP、SSP、メディアを立ててプログラマティック広告の実環境に近い状況でOPが正常に流通するかを実験した。



3 社会実装のイメージ（具体的なユースケース、想定ユーザ）

導入先	詳細	ペインポイント	解決できる課題
メディア	新聞・雑誌・放送局などのWebサイト	●偽サイトを作成され、詐欺的行為などに悪用される ●アテンションエコノミー型MFAサイトへの広告費の流出	●偽サイトによる信用毀損の防止 ●不正サイトへの広告費流出の防止
自治体などの公共団体	政府機関・自治体・警察・医療機関などのWebサイト	●偽サイトを作成され、詐欺的行為などに悪用される ●災害時にSNSなどで偽・誤情報が発信され、災害対策のや市民生活に混乱が生じる	●住民の詐欺被害防止 ●社会混乱の予防
企業	金融機関・生活インフラ企業などのWebサイト	●類似ドメインを利用して偽サイトやフィッシングメールを作成され、顧客が口座情報や個人情報を漏らしてしまい、詐欺的行為などに悪用される	●ユーザーの被害防止と、詐欺不安によるサービス利用低迷の予防
広告主	広告主が出稿するデジタル広告およびそのWebサイト	●アドフラウドにより実際に掲載されなかったり、悪質なサイトに広告が掲出され、ブランドイメージが毀損する ●偽広告が氾濫し、広告そのものへの信頼が損なわれる	●ブランドイメージが守られる ●出稿した広告への信頼が得られる
プラットフォーム	検索エンジンやSNSなどのプラットフォームが提供するWebサービス	●検索結果画面に詐欺サイトがリンクされ、ユーザーが詐欺被害に遭う ●SNSコンテンツの信頼性を示す客観的な手段がない	●利用者離れの防止 ●信頼できる情報の提供
著名人	芸能文化スポーツ領域で活動する著名人	●肖像を無断利用した詐欺広告が掲出され、消費者被害が発生するほか、自らのイメージが損なわれる	●自らのイメージ保護とファンの詐欺被害防止
ネットユーザー	生活の中でインターネットを利用し、各種情報に接触する人々	●偽・誤情報を正しく見分けることができない ●インターネットを悪用したサイト、広告等による詐欺被害にあう ●SNSなどで拡散してしまい、意図せず社会不安の原因を作ってしまう	●正しい情報を見分け、安心してインターネットを利用できる環境を得る ●自己の情報を安心して利用するサービスに入力できる ●情報を拡散する前に、信頼できる情報なのかを確認しやすくなる

ORIGINATOR PROFILE

株式会社Classroom Adventure

偽・誤情報サンドボックスを活用した実践的ゲーム型プレバンキング技術の開発・実証

1 開発・実証の背景・目的

生成AIの普及により偽・誤情報が高精度化・低コスト化し、短時間で大量に作成され同時多発的に流通し得る状況が生じている。
この環境下では、個別の訂正情報を提示するだけでは限界があり、情報の受け手一人ひとりが、未知の事例に対しても疑い立ち止まり確かめ拡散しないという判断行動まで含む実践的リテラシーを継続的に実行できる状態が求められる。
そこで本事業では、偽・誤情報の「作られ方」と「もっともらしさの構造」を安全に体験できる偽・誤情報サンドボックスと、反復利用を前提としたゲーム型プレバンキング学習設計を開発・実証し、受け手の判断行動を社会に広げることを目的とする。

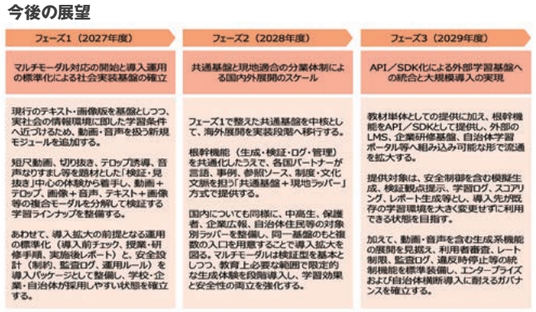
2 開発技術の概要

偽・誤情報を“作る側”として体験し「なぜもっともらしく見えるのか」を構造理解したうえで 確認行動へ転換する」実践的ゲーム型プレバンキング技術を開発。
中核は 偽・誤情報サンドボックスと 反復学習を前提とした学習設計であり、学校や自治体等の現場で実施可能な運用基盤まで含めて整備。

3 社会実装のイメージ

学校・自治体・企業等の現場で「単発イベント」ではなく反復運用 継続利用できる形で提供することを想定。
導入側が「実施できる 予算化できる リスク管理できる」と判断できる導入パッケージ（手順・体制・必要機材・安全配慮）を整備し、普及導線を設計する。

今後の展望



CLASSROOM ADVENTURE
Tokyo-based Gamification Team

株式会社コンステラセキュリティジャパン

偽・誤情報およびデバンキング情報拡散のシミュレーション技術の開発・実証

1 開発・実証の背景・目的

解決したい課題	本技術におけるソリューション
生成AIによる脅威の増大と対策リソースの非対称性	偽・誤情報やデバンキング情報の拡散予測および拡散シミュレーションにより対策リソースの分配を最適化
日本市場・偽情報対策に特化した分析ツールの不在	偽・誤情報分析の実務経験を持つ日本語話者の知見を反映した分析ツールの開発
国際的な情報共有の潮流と国内における連携不足	データをSTIX形式で構造化することによりデータ共有の円滑化を実現
変化する情報環境と既存ソリューションの限界	モジュール式アーキテクチャの採用により機能追加や更新への柔軟性を確保

2 プラットフォームアーキテクチャ

① OpenCTI	国際的な脅威情報共有の枠組みであるSTIX形式対応のプラットフォーム
② コネクター機構	収集および加工したオープンソースデータをSTIX形式に変換し、OpenCTIに格納
③ エンリッチメント機構	データを自然言語処理等により加工し、傾向分析を実施
④ 分析機構	取得および加工したデータに対し、より高度な分析を実施

3 社会実装のユースケース：SNS 上の情報拡散の分析

政府機関	海外勢力による議論介入の特定、および対策のための起点の特定
	戦略的コミュニケーションの設計
報道機関	偽・誤情報拡散の影響度の推測
	デバンキング報道の有効性の評価
民間企業	顧客のセグメンテーション
	マーケティング効果予測・測定

CONSTELLA SECURITY JAPAN

株式会社TDAI Lab

デジタル情報空間における多層的意味解析と拡散ダイナミクス解明プラットフォームの開発・実証

1 ミスリーディング情報の意味解析

■ SNS上では、短文で感情を揺さぶるミスリーディングな表現（論理的誤謬）が増加しており、従来の真偽判別だけでは対応不可能

■ 本開発・実証では、大規模言語モデルによる深層的意味解釈エンジンによりSNS投稿内の論理的誤謬（14種類）やナラティブを解析し、情報拡散分析エンジンで結果をレポートとして統合

ミスリーディング情報の構築論理・影響を包括的に分析・可視化

2 分析レポート出力

■ 分析目的に応じて3種類のレポートを自動作成するLLMエージェントを構築
- ✔ **SNSデータを用意せずとも利用可能**
- ✔ **4000件のSNSデータを1時間以内で処理可能**

■ 2025年総裁選に関するYouTubeコメント1500件を本システムで分析したレポート出力例:

3 ユーザーヒアリング結果

■ マスメディア、調査・分析機関等の有識者10名にヒアリングをし、分析レポートの有用性を評価

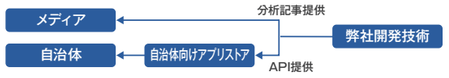
■ 迅速性および理解容易性は比較的高く評価された一方、信頼性および業務接続性は課題として顕在化

高評価	低評価
・リファレンスを元にファクトを整理してくれる ・話題やナラティブなど様々な角度で定量的に分析できる ・読み物として面白い	・分析手法の説明を書いてほしい ・内部のプロンプトを編集したい ・どこまで事実ベースで、どこから推論ベースなのか分かりにくい

4 社会実装実績

■ 社会的インパクトの大きい4テーマについて、本技術を利用した分析記事をメディアへ寄稿
1. 「**JICAがホームタウン撤回 SNS炎上招いた「私は被害者」**」（2025年9月29日公開）
2. 「**高市氏、SNSでは敵無し「愛国」と「感情」の支持投稿**」（2025年10月15日公開）
3. 「**「国も動いたクマ騒動 AI動画が描く「もうひとつの世界」**」（2025年11月18日公開）
4. 「**SNS、自民党嫌いのサナエ推し ファンが支える高市政権**」（2026年2月16日公開）

■ プラットフォーマーとのAPI連携による自治体への提供



TDAI Lab

株式会社データグリッド

SNS ユーザー支援を中核とした偽・誤情報対策の開発・実証

1 開発・実証の背景・目的

偽・誤情報がSNS上で生産・拡散・消費される中で、本質的な問題解決には、メディアや官公庁、企業での対策だけでなく、SNSユーザーが自律的に偽・誤情報に対処するリテラシーやスキル、行動習慣を持つこと重要である

事前（SNSユーザー）	拡散（SNSユーザー）	訂正（メディア、官公庁、企業）
リテラシーの向上	適切な情報検証 生成AIへの対応	真偽不明言説の収集 訂正情報の周知

2 開発技術の概要

マルチモーダル対応の検知技術などの5つの要素技術を開発

①	②	③	④	⑤
ディープフェイク検出の運用・高度化技術	信頼性評価付き根拠情報検索技術	画像・動画に対する出所および文脈検証技術	情報摂取・行動の分析技術	リスクアセスメント・拡散予測技術

3 社会実装のイメージ

SNSユーザー向けのファクトチェック支援アプリを中心に展開することで、情報リテラシーの向上に貢献し、報道機関や官公庁での利用を想定したプロフェッショナル版も展開する

DATAGRID

関西テレビソフトウェア株式会社

放送波を活用した災害時における偽・誤情報対策技術の開発・実証

1 防災コンパスの開発

災害時などの通信不安定な状況でも「次の一歩」の判断を支援する、コンパス型アプリケーションの開発。

2 FA コードの開発

拡散された画像について、ファクトチェック判定の有無を確認することができるFAコードの開発。

FAコード（Fact Annotation Code）
本実証において定義した独自のファクト注釈識別子

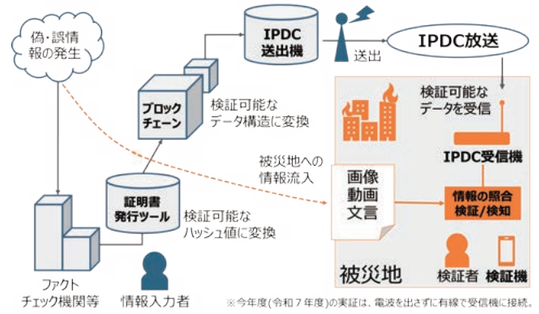
3 開発・実証の背景・目的

災害時では、情報の空白地帯を縫うように様々な偽・誤情報が拡散されてしまい、避難所で暮らす被災者や災害時支援に携わる人々を困惑させています。

そこで、地上波デジタル放送という日常的に使用されている災害にも強いインフラ活用し、偽・誤情報に惑わされないための「信頼できるデータ」を、地域の人に効率よく周知する方法を考案しました。

4 社会実装のイメージ

広域配信に強みのある放送に、IoT機器など制御可能なデジタルデータを重畳します。
ブロックチェーン技術を活用することで、高い信頼性を維持することができます。



※今年度（令和７年度）の実証は、電波を出さずに有線で受信機に接続。

関西テレビソフトウェア株式会社

SEARCHLIGHT株式会社

ストリーミング動画コンテンツの真偽検証支援ツールの開発・実証

1 開発・事象の背景・目的

本事業は、ストリーミング動画プラットフォームの閉鎖性や検証作業の属人性に起因する「検証コストの高さ」を解消し、ステークホルダーが直面するリソース不足やブランド毀損リスクを低減することを目的とします。

また、AIによる自動判定ではなく、「主張の構造化」と「出典の可視化」を通じた判断支援を行うことで、検閲リスクを回避しつつ情報の透明性を高める新たな社会的基盤の構築を目指します。

2 開発技術の概要

本開発では、影響力を増すストリーミング動画の解析を目的として、「動画内容の構造化」と「根拠情報の提示」を行う真偽検証支援技術の開発・実証を行いました。

具体的には、大規模言語モデル（LLM）をチューニングし、動画内の主張とその検証のためのエビデンスを論拠強度を加味したうえで収集し、自動照合するシステムを構築いたしました。

3 社会実装のイメージ

新聞社や報道機関などのマスメディアや、ファクトチェック団体などのプロフェッショナル向けの提供を予定しております。なお、実証期間中においては、2025年7月の参院選、2026年2月の衆院選において想定ユーザーへの試験提供を行いました。

また、本事業で開発した技術を応用し、ストリーミング動画コンテンツの法令適合性のチェックや、炎上監視を含むモニタリングも予定しております。

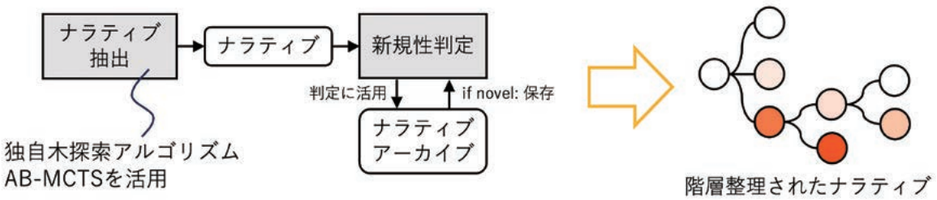
SEARCHLIGHT

Sakana AI株式会社

画像・動画を中心としたSNS上の投稿の真偽判定システムの開発・実証

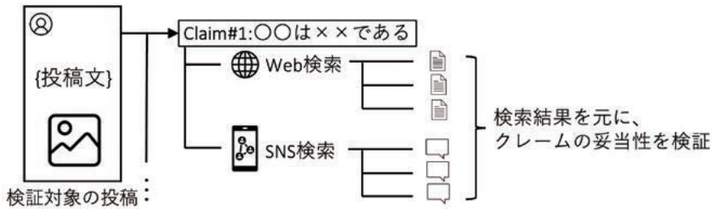
1 SNS 空間の可視化

ナラティブ（言及内容）をAIにより抽出・クラスタリングし、ツリー形式で整理



2 総合的な偽情報判定

動画・画像・インターネット・SNS・反応など、多面的な観点をAIが自動分析



3 対策案の立案

AIを用いたSNS空間シミュレーションにより、効果的な打ち手（発信）を立案

sakana.ai

サン電子株式会社

多元統合型偽・誤情報検出技術の開発・実証

1 生成 AI で偽情報が日常化：揺らぐ社会の「信頼」

生成AIの普及により、誰でも高品質な画像・動画・音声を偽造できる時代となり、偽・誤情報の流通量が加速度的かつ指数関数的に増大している。
行政判断や報道、企業活動、個人の行動選択まで誤誘導され、社会の意思決定の前提が揺らぐだけでなく、経済安全保障上の脅威や不信・混乱の拡大が現実的なリスクに。

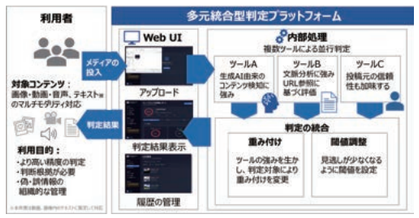
2 世界は対策を加速中：規制 × 技術の多層アプローチ

海外では、規制強化やファクトチェック、技術標準化（透かし等）などの多層的な対策が進展。一方で、単一技術への依存や運用の属人化といった課題は残り、実務で使える「判断基盤」「運用ルール」は依然として不足している。

3 “断定”より“判断支援”：日本社会に適合した偽誤情報対策プラットフォーム

本事業は、完全自動での真偽判定における課題（誤検知や根拠不足）を補完して、現場の意思決定を支える「判断支援プラットフォーム」の構築を目指した。
複数ベンダーの技術を統合して画像・動画・音声を一元評価。さらに運用において真／要検証／偽の三段階判定で情報の優先度を明確化し、メディアの信頼性判断、企業の風評被害対策、公的機関の情報環境保全を支援する。
また、既存システムへの容易な組込を可能にしSNSプラットフォーマーの基盤強化にも貢献する。

4 複数判定を束ねて精度を上げる：多元統合型・偽誤情報検出技術



現場の「意思決定」を支える3つのメリット

①**統合判定による高精度化**:複数技術を動的に重み付け統合し、単独ツールでは検知困難な高度な偽造も捕捉する。

②**根拠に基づく判断支援**:スコアに加え参照URL等の根拠を明示。担当者が自信を持って判断できる材料を提供する。

③**組織的な履歴管理**:判定や対応の履歴をダッシュボードで一元管理し、ガバナンス強化とナレッジ蓄積を実現する。

サン電子株式会社

NABLAS株式会社およびNTT東日本株式会社

電話音声フェイク検知および自治体向け偽・誤情報総合対策の開発・実証

1 AI のなりすまし・詐欺から守る安心のインフラ技術
生成 AI 時代の電話音声フェイクを見破る検知技術

電話環境下でのフェイク音声をAIが検知しアラートでお知らせ

活用が期待される場所：通信事業者／Web会議プラットフォーマー

拡大する脅威　現実の被害として発生中！
- AIによる声の完全模倣
- 上司を装う送金詐欺

開発のポイント
- 環境音やノイズがあっても検知可能！
- 固定・携帯・IPまで幅広い環境で検知可能！

NTT東日本との共同開発により、実環境で機能する検知アプリを実現

2 「検知」と「証明」で情報の信頼性を確保
Web・SNS 上のフェイクに惑わされない情報受発信技術


実際の検知画面

※画像が小さいため、詳細はQRコードよりご確認いただけます。

⚠ 生成AIのフェイク情報による混乱が増加中！

災害時のデマ拡散／クマ情報のデマ拡散／選挙時のデマ拡散

SNSで情報収集する官公庁・企業が混乱

官公庁・企業の公式情報の信頼性低下

技術 ① 総合的フェイク検知技術

SNS上の画像・動画・テキスト情報を複合的に解析しフェイクを見破る

SNS上のフェイク情報を高精度に検知

一部加工や改変も検知

技術 ② 真正性証明（電子透かし）

正式発信者からの情報にウォーターマークを埋め込み「正しい情報」を保証

「本物の証」である電子透かしを付与

市民が信頼できる情報を判断できる仕組みを構築

3 音声フェイク検知、自治体向け偽・誤情報総合対策の双方で社会実装にこだわった開発
「実際に使える」技術を見据えた実証実験の実施

実運用への適応性・社会実装を重視した「使える」技術にするための実証実験

通信インフラ実環境での検証（音声フェイク検知）

NTT東日本の実際のひかり回線ルートを使用し検証

ひかり電話／携帯電話／IP電話

コーデック変換・エコーキャンセル処理を含む実環境での検証

自治体実運用を想定した検証（偽・誤情報総合対策）

実証フィールドとして長野県伊那市と協力

SNSへの投稿 → 伊那市

偽・誤情報対策システム
- フェイク検知
- 電子透かし・改ざんチェック

情報の受発信の実務フローに合わせた検証

NABLAS　NTT EAST

日本電気株式会社

請負契約件名：
AIを活用した情報コンテンツの真偽判別支援技術の開発・実証

1 開発実証の背景・目的

偽・誤情報による国民の混乱が減少する社会の実現に貢献

インターネット上に氾濫している偽・誤情報により、社会が不安定化し、混乱リスクが高まる中、実際に問題となった事例も年々増加中。とくに、信頼性の高い情報を発信する責任のあるマスメディア（放送局等）や自治体等では、偽・誤情報の介入を防ぐ対策に多くの労力を割き、業務負荷が高まっている状況。

その社会課題解決に対し、弊社では、テキスト・動画・画像・音声からなる複合コンテンツの真偽判別支援技術を昨年度開発。今年度は、社会実装を推進すべく、昨年度得た課題に対する機能強化・高度化を進めると共に、ユーザー先となりうるマスメディアや自治体等へのヒアリングを継続して実施。

2 開発技術の概要

昨年度開発成果をベースに「判別率向上」「使う人を選ばないUIと分析プラン」を実現

開発内容

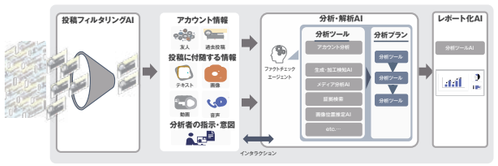
SNSから偽情報疑いのある情報をフィルタし、情報コンテンツや分析者の意図に応じて適切に分析・レポートするシステムとして主に以下のコンポーネントを開発。

- 投稿フィルタリングAI
- ファクトチェックエージェント
- 各種分析ツール（画像位置推定、SNSアカウント分析、証拠検索、生成・加工検知など）

性　能

令和6年度事業で開発した偽・誤情報判別支援システムと比較して誤判別率を80%低減（判別率向上）

SNSの投稿　　偽・誤情報総合分析支援システム



3 社会実装のイメージ

マスメディア・自治体をメインとした段階的導入によるターゲットの拡大とクラウドサービス基盤を軸とした展開を予定

ユーザ先としては、マスメディアと自治体をメインに、今後更なるマーケットでの利用拡大を検討。

具体的なユースケースとして、マスメディアが番組制作で利用するSNSや視聴者投稿など多岐にわたるインターネット情報の真偽判定を、AIが自動化・効率化することで、人力による膨大な工数を削減。

また自治体では災害時、SNS上の偽情報・誤報で自治体公式情報が埋もれる課題を本技術で検知・判定し、自治体の迅速な正確情報発信を支援。

クラウド基盤のサービス型提供により全国的な導入を促進し、将来的にはBtoBtoCモデルにより、国民利用まで視野に入れた発展を目指す。

NEC

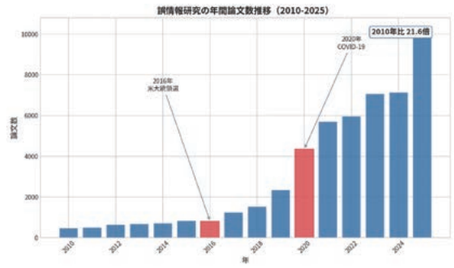
株式会社新領域安全保障研究所

インターネット上の偽・誤情報等への対策技術の開発・実証事業

グローバル・メタアナリシスと国内実証による対策技術の有効性の研究・調査

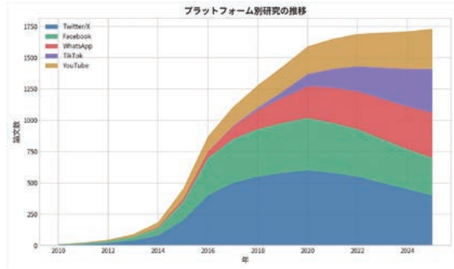
1 15 年間で論文数は 21.6 倍に

- OpenAlexで抽出可能な誤情報・偽情報に関する学術研究は、2010年の368件から2024年には7,950件へと急増し、その規模は21.6倍に（右図）
- 2016年の米大統領選後の政治的プロパガンダへの懸念や2020年のCOVID-19で特に増加か。昨今ではAIをキーワードに含む学術研究も増加。



2 対象プラットフォームの傾向

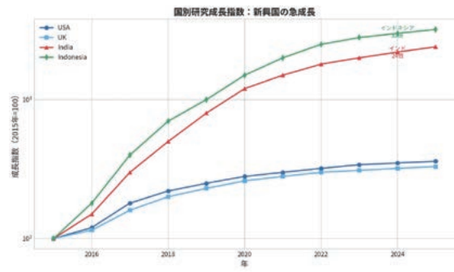
- データアクセスの容易さから、Twitterを対象とした研究が目立ったが、最近はAPI料金の高騰の影響もあるのか減少傾向（右図）
- 近年はTikTok/YouTubeに関する研究も増加



3 新興国でも論文が急増

- 英米に比較して、インドやインドネシアの研究機関に所属する著者の論文が激増（右図:対数グラフ）
- 絶対量（2010-2025）では、米国と英国だけで全体の約40%でまだまだ英米の研究が目立つ。

注:英語圏以外の論文は正しく抽出できていない可能性が高く、実態はもっと多い可能性もある。



INODS

東京大学大学院情報学環

生成AI時代における偽誤情報流通と認知特性の解明に関する研究・調査

1 研究・調査の背景と概要

調査・研究の背景

- 生成AIによる偽誤情報の**高度化**と**情報過多**
- 人間の認知能力の限界が課題
- 技術だけではなく、人の判断プロセスを踏まえた対策が不可欠

調査・研究の概要

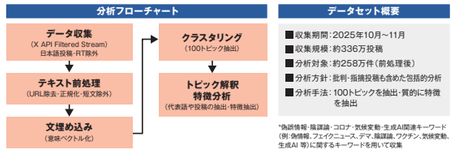
日本語偽誤情報流通をデータから体系的に理解し、人間中心の対策モデル構築を目指す

研究項目1 日本語偽誤情報脅威の体系化

研究項目2 生成AI偽誤情報対策警告介入の効果の実証

2 研究1 研究・調査手法

- X（旧Twitter）APIを用い、偽誤情報と生成AI関連の日本語投稿を約336万件（リツイートを除く）を収集*・分析
- トピックモデル等を用い、偽誤情報言説の特徴を構造化・体系化



3 【研究 1】結果

- メディア・政治不信への言及が最多、次いで誤情報を指摘、排外主義的が続く
- 情報操作の手法枠組（DEPICT）による分析の結果、**信用失墜型（Discrediting）**が多い→発信元を攻撃することで正しい情報さえも**受け付けない土壌**が形成されている懸念

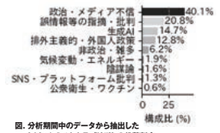


図. 分析期間中のデータから抽出したトピックのメタカテゴリごとの投稿割合

分類	戦略 Strategy	手法 Device	対象 Target	具体的フレーズの例
メディア批判	信用失墜	なし	マスコミ	「テレビ局は不祥事を報道せず隠蔽した」
	信用失墜	なし	なし	「いじめ隠蔽に専われていく」
	信用失墜	疑問・疑念	マスコミ	「偏向報道があまりにも酷すぎた（関しかけ）」
	挑発・煽り	なし	なし	（感情的な攻撃や嘲笑）
政治政策批判	信用失墜	なし	特定の個人・団体	「クソデマ」「帯状ですよ？」

表. 日本語言説空間でよく見られる手法の例

4 研究2 研究・調査手法

- 日米独仏印の5カ国でオンライン実験を実施（N=5,443）
- **強い警告（メッセージ型）**と**控えめな警告（ラベル型）**の効果を検証
- 動画の正確性評価*に加え、ユーザーの迷い（不確実性）を測定
- 個人属性や地域、AIの認識、性格、価値観などで制御の上比較

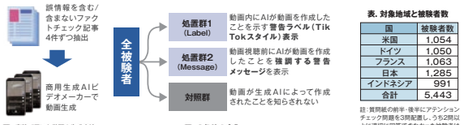


図. 実験で用いた動画の生成方法

図. 3条件の介入

表. 対象地域と被験者数

国	被験者数
米国	1,054
ドイツ	1,050
フランス	1,063
日本	1,285
インドネシア	991
合計	5,443

註：被験者は各動画を視聴後内容の正確さを7段階のリッカート尺度で選択（あなたの知る限り、動画の内容は正確ですか？）

註：質問紙の前半・後半にアテンションチェック問題を3問配置し、うち2問以上に適切に回答できなかった被験者は対象被験者から除外、表中被験者数は除外後の数

5 【研究 2】結果

- **強い警告**：動画への信頼度（正確性評価）をを下げる効果
- **控えめラベル**：信頼度は下がらずどちらとも言えない（判断の迷い）が増える
- 識別能力の限界：警告は全体の警戒心を高めるが本物と偽物を正しく見分ける力を向上させる効果はない
- 警告のデザイン以上に、個人の価値観、AIへの態度や陰謀論親和性が判断を左右する

警告の強さ	表示イメージ	反応の傾向
強い（メッセージ型）	目立つ字体で繰り返し警告	正確性評価がやや下がる
弱い（ラベル型）	隅っこに小さな文字	「正しいとも誤っているとも言えない」が増える

表. 主な結果

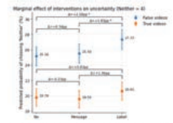


図. 介入が被験者の迷い（不確実性）に与える影響

6 結果・考察・期待される効果

- 単なる真偽判定ではなく「人がどう受け取り、どう判断し、どう行動するか」から認知と信頼を守るための設計が必要
- 現状の主流である控えめなAIラベルだけでは、誤情報対策として不十分な可能性
- 技術的対策に加え、ユーザー自身の認知バイアスへ働きかける教育が不可欠
- 現状把握や将来の脅威の理解を深めるためにも、SNSプラットフォームにおけるデータ公開・透明性を求める必要性

東京大学大学院 情報学環
澁谷研究室

偽・誤情報対策技術でできること①

あなたはどの立場ですか?

- 公的機関
- 企　業
- 一般市民

偽・誤情報対策技術でできること②

公的機関の皆さま

想定リスク

- 災害・選挙等における偽・誤情報拡散
- 公式発表の切り取り・誤引用
- 初動対応の遅れ

凡例:
- 技術区分Ⅰ コンテンツの真偽判別支援・改ざん検知技術
- 技術区分Ⅱ 真正性保証・信頼性判断支援技術
- 技術区分Ⅲ 情報流通情報の可視化・分析技術
- 技術区分Ⅳ 情報の拡散防止・無効化技術
- 本事業で注力している取組対象

偽・誤情報対策技術でできること

技術区分Ⅰ 拡散中の動画等が実際の被害状況を撮影したものか真贋検証することにより、改ざんされた情報に基づく誤った初動判断を防止できる

技術区分Ⅱ 行政が発信した正規の情報であることを示す真正性情報を付与・表示し、住民が一目で公式情報と判断できる

技術区分Ⅲ 自動モニタリングツールとして活用することで、夜間・休日に急増する投稿等の情報を自動で収集し、担当者の張り付き負担を軽減できる

技術区分Ⅳ SNS上で急増している関連投稿に対し、投稿閲覧時点で注意喚起を表示させることで、利用者に慎重な判断を促すことができる

偽・誤情報対策技術でできること③

企業の皆さま

想定リスク

- ブランド毀損・株価低下
- なりすまし詐欺被害・偽サイトへの誘導
- 拡散中の訂正判断困難

凡例:
- 技術区分Ⅰ コンテンツの真偽判別支援・改ざん検知技術
- 技術区分Ⅱ 真正性保証・信頼性判断支援技術
- 技術区分Ⅲ 情報流通情報の可視化・分析技術
- 技術区分Ⅳ 情報の拡散防止・無効化技術
- 本事業で注力している取組対象

偽・誤情報対策技術でできること

技術区分Ⅰ 投稿内容が実際の出来事に基づくものか、過去素材の流用や改ざんはないかを確認することにより、速やかにブランド価値や株価の低下の防止に向けた初動対応ができる

技術区分Ⅱ 自社が正式に発信したコンテンツであることを示す真正性情報を付与し、受け手に公式性を可視することで誤認を予防できる

技術区分Ⅲ 投稿を論点別に自動整理し、論点ごとの拡散量・誤解が生じやすい表現のパターンを並べて提示することで、「どの論点を優先して説明するか」の根拠整理を支援できる

技術区分Ⅳ 偽・誤情報の手口や真贋検証の手法に関する学習コンテンツを活用することで、偽・誤情報を適切に判別できる

偽・誤情報対策技術でできること④

一般市民の皆さま

想定リスク

偽・誤情報を信じてしまう

偽・誤情報に気づかぬまま拡散してしまう

なりすまし詐欺被害・偽サイトへの誘導

凡例:
- 技術区分Ⅰ コンテンツの真偽判別支援・改ざん検知技術
- 技術区分Ⅱ 真正性保証・信頼性判断支援技術
- 技術区分Ⅲ 情報流通情報の可視化・分析技術
- 技術区分Ⅳ 情報の拡散防止・無効化技術
- 本事業で注力している取組対象

偽・誤情報対策技術でできること

技術区分Ⅰ SNS投稿内容が事実に基づくものかを確認することにより、誤った情報を無意識に拡散してしまうリスクを低減できる

技術区分Ⅱ 広告画像やWeb表示が正規の事業者によるものか確認することにより、偽ECサイトや投資詐欺への誘導を回避できる

技術区分Ⅲ 大量投稿を内容の近さで自動グルーピングし、「どの話題がどれくらい増えているか」を主要論点に圧縮して表示することで、状況理解を支援できる

技術区分Ⅳ 共有ボタンを押下時に「未確認情報の可能性」等と表示することで、利用者が共有前に事実確認を行う機会を設けることができる

代表的な被害事例のご紹介①

✔災害発生時、SNS上では過去に別地域で発生した被害映像や虚偽の被害情報が拡散する事例が見られます。正規の報道映像を装った投稿もあり、報道機関の公式映像と誤認されることで、誤情報に基づく救助要請や救助隊の誤出動が発生したケースも確認されています。

発生した事象

無関係な映像の流用・拡散

過去の別の災害の津波映像を現在進行形で発生している事象として投稿。精緻な情報（日時・場所）の記載はなく、映像による視覚的インパクトのみで「証拠映像」として拡散

虚偽の救助要請

実在しない住所や、事実に基づかない架空の被害状況を記載した「救助要請」がSNSで氾濫。中には善意のユーザーによる拡散もあり、情報の真偽判定が困難な状況に陥る

拡散を促進した要因

情報の空白期

災害発生直後は公的機関や報道機関の正確な情報の確認と発信に時間を要する。この情報の空白時間を埋める形でリアルタイム性に優れたSNS投稿が先行

情報利用者層形成期

注目や反応数を得ることを目的に、刺激的な偽映像や虚偽の救助要請を投稿するアカウントが出現

速報性優位フェーズ

一般ユーザーは情報の真偽よりも速報性を優先する傾向にある。 SNSは反応の多い投稿ほど目に触れやすくなる仕組みがあるため、内容を信じた利用者による再投稿が連鎖し情報が爆発的に拡散

リスクと対応技術

■災害時の偽・誤情報による判断ミス

未確認の津波映像や救助情報を信じることで避難が遅れる/不要な出動が発生するなど生命・救助資源に直結する判断ミスが生じてしまう可能性がある

■公式情報の誤認による混乱

SNS上の情報を公式発表と誤認し、企業や組織内で誤った前提で意思決定・連絡が行われてしまう可能性がある

■初動対応の遅れによる市民の混乱

自治体・報道機関が迅速に正確な情報を発信できず、危機広報が遅延し市民の不安が増幅してしまうおそれがある

対応技術

- 複数コンテンツの情報を照合し、多面的に分析することにより真偽見極め、改ざん検知を支援 → 技術区分Ⅰ
- 電子署名や電子透かしを用いて真正性を証明 → 技術区分Ⅱ
- 情報流通空間全体のトレンドを可視化・分析 → 技術区分Ⅲ
- 拡散範囲の制限・改ざん疑いのあるコンテンツへのタグ付けにより、情報の受け手へ注意喚起 → 技術区分Ⅳ

代表的な被害事例のご紹介②

✔偽情報は、インターネット空間を通じて急速に拡散し投資家の不安を連鎖的に増幅させます。その結果、市場価格や企業の信頼性そのものに影響を与えてしまうケースが多数報告されています。企業は、拡散の抑止と信頼性確保を両立する技術的対策が不可欠です。

発生した事象

偽プレリリースによる株価急落

CEOの辞任・謝罪を模した偽のAI音声や偽プレリリースが拡散。市場が真に受け、公式否定まで株価が混乱

偽の工場爆発動画の拡散

実在企業の工場爆発や国家主要施設の火災等、「AI生成動画」がニュース風アカウントにより拡散し、株価の下落が発生

偽公式SNSによるリコール連絡

大手企業を模した認証バッジ付きの偽企業アカウントが商品への異物混入による商品のリコールを発表。投資家間に限らず、一般市民の消費行動も混乱

拡散を促進した要因

生成・信頼性の演出

生成AIにより、本物に酷似したロゴ・書式・文体を模倣。過去のIR文書学習や認証バッジ悪用により一目では真偽判別困難な状態を演出

デジタル空間での連鎖的な不安増幅

人的措置の介入前に、投資家掲示板・まとめサイトなどへの転載に加え、SNSアルゴリズムを介して連鎖的に投資家らの不安をあおり株価急落

市場反応・二次拡散

株価急落そのものが「何があったのか」という疑念を呼び報道機関等がさらに拡散。企業が訂正対応を実施したとしても、当該情報は拡散速度で劣後し、回復に時間を要する

リスクと対応技術

■市場価格の急変動

偽情報を真に受けた投資家の売りが集中して実態と無関係に株価が急落。企業価値が毀損されることに加え、指数構成銘柄への波及や変動幅上昇による金融市場全体へ影響が生じてしまうおそれがある

■信頼性の低下

既存顧客との契約見直しが発生したり、既存の公式SNS等の発信チャネルの信頼が失墜してしまうおそれがある

■風評被害

偽情報が企業と無関係な主張の拡散や便乗商法に使われ、ブランド価値が毀損されてしまうおそれがある

対応技術

- インターネット上の情報等、複数のコンテンツを組み合わせてディープフェイクを見破る
  → 技術区分Ⅰ
- 電子署名・電子透かしを用いて真正性を証明
  → 技術区分Ⅱ
- 偽情報が拡散しそうなトピックや被害に遭いそうな状況を事前に予測
  → 技術区分Ⅲ
- 偽の情報が流れる可能性を事前に注意喚起
  → 技術区分Ⅳ

MEMO