日本総研
The Japan Research Institute, Limited

ディープフェイク検出技術
～現状と金融取引の防御に向けた課題～

2026年11月12日
株式会社日本総合研究所
先端技術ラボ

執筆者：先端技術ラボ 渡邊 大喜、會田 拓海、稲田 龍人
技術調査協力：日本電気株式会社 セキュアシステムプラットフォーム研究所
＜お問い合わせ＞ 当社ホームページの問い合わせフォームよりご連絡ください。

- 本資料は作成日時点で弊社が一般に信頼できると思われる資料に基づいて作成されたものですが、情報の正確性・完全性を保証するものではありません。本資料の内容は、経済情勢などの変化により変更されることがあります。本資料の情報に起因して閲覧者及び第三者に損害が生じた場合も、執筆者、取材先及び弊社は一切責任を負いかねます。
- 本資料の著作権は株式会社日本総合研究所に帰属します。本資料の一部または全部を、電子的または機械的な手段を問わず、無断で複製または転送などを行うことを禁止しています。

Copyright (c) The Japan Research Institute, Limited

はじめに

生成AIの高度化により、実在の人物と見分けのつかない画像・音声・動画を容易に生成できるようになった。こうしたディープフェイクは、偽情報の拡散にとどまらず、人物のなりすましによる送金詐欺、本人確認の突破による不正口座作成、保険請求等における証拠の偽造など、金融犯罪にも用いられている。

対抗手段となるディープフェイク検出技術は、生成時に透かしや来歴情報を付与する能動的手法と、生成物そのものを分析する受動的手法の双方で研究・実用化が進んでいる。一方、未知の生成器や実環境下での精度低下、検出回避攻撃への脆弱性など、検出技術単独による防御には限界がある。

本レポートでは、ディープフェイクの社会的影響と検出技術の全体像を整理し、技術・企業・規制の動向を概観する。さらに、金融取引の防御に向けた今後の課題と提言を示す。

INDEX

1. ディープフェイクの社会的影響
2. ディープフェイク検出技術の概要
3. 検出技術の最新動向
4. 社会実装・規制の動向
5. 今後の課題と提言

※本レポートの執筆にあたっては、ディープフェイク検知や偽・誤情報分析に豊富な知見を有する **日本電気株式会社 セキュアシステムプラットフォーム研究所　AIトラスト研究チーム(P.24)** より、多大なご協力をいただきました。

Copyright (c) The Japan Research Institute, Limited

エグゼクティブサマリ

- ディープフェイクは既存の金融犯罪に組み込まれ、攻撃の手口を巧妙化・大規模化している
  - ディープフェイク犯罪は、偽られる対象によって「人物・企業」「本人性」「損害・請求証拠」の3類型に整理できる

- 検出技術に関する研究は、画像・音声・動画など各モダリティ全般で加速。**近年はマルチモーダルな検出手法（映像と音声など複数のモダリティを組み合わせて判定する）に関する研究**も増えている

- 技術開発は、「真正性の活用」・「説明可能性」・「汎化」・「マルチモーダル活用」の4つの方向で進展している
  - 単独の検出器による二値判定から、**複数の証拠に基づき、真正性と判断根拠を示す検証**へと移行しつつある

- 社会実装では、海外は金融・政府で検出技術の導入が進み、国内は公的事業のもとで開発・実証が進展

- 規制面は、各国でAI生成コンテンツの表示義務の施行が相次ぎ、生成AI大手は透かし・来歴の付与を進める

- 検出技術単独での防御は困難であり、制度的基盤のもとで**技術と業務手順を組み合わせた多層的な対策**が必要
  - 検出技術には「汎化性能」「実環境への適応」「回避攻撃への耐性」「説明の信頼性」の課題が残る

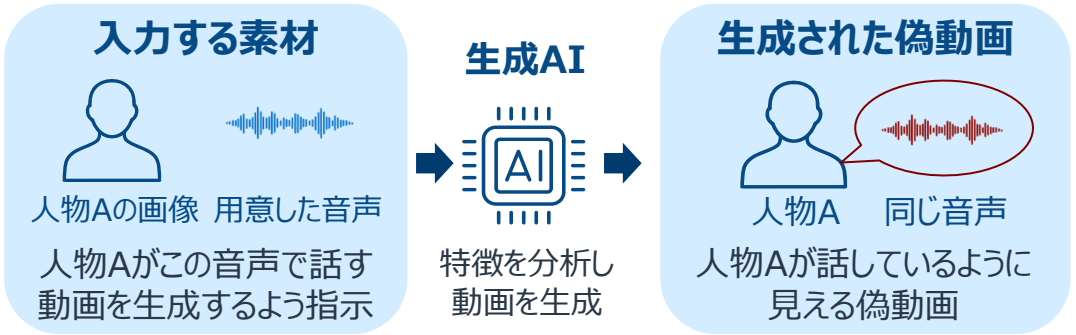
Copyright (c) The Japan Research Institute, Limited

日本総研
The Japan Research Institute, Limited

ディープフェイクとは

- ディープフェイクとは、ディープラーニング（深層学習）とフェイク（偽）を掛け合わせた造語
- 生成AI等を用いて、実在する人物や出来事を本物らしく再現・改変したコンテンツ、またはその生成技術を指す

ディープフェイク概要と悪用の事例



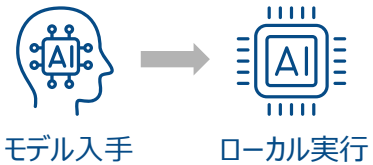
性的ディープフェイク	SNS上に投稿された顔写真などを用いて、偽の性的画像を生成する事例が確認されている*1,2
広告利用	著名人の肖像・音声をディープフェイクで無断利用した、なりすまし詐欺広告が流通している*3
政治	ロシアのウクライナ侵攻で、ゼレンスキー大統領が降伏を呼びかける偽動画が拡散された*4
愉快犯・悪ふざけ	街中にクマが出没したとする偽画像・偽動画が複数投稿された*5 2026年の熊本地震で、被災者の表情やポーズを変えた偽画像が作成・拡散された*6

ディープフェイクの生成方法の例

AIサービスは規約違反のコンテンツ生成を制限しているが、安全制御を回避して生成する例もある。また、規約内で生成したコンテンツでも、利用・拡散の仕方によっては社会混乱につながる

1 ローカル環境の生成AIで生成

生成制限のないモデルを選び、手元の環境で実行
配布元の例：Hugging Face、GitHub



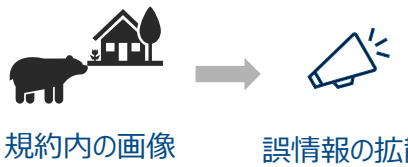
2 AIサービスの安全制御を回避して生成

入力の工夫でAIサービスの安全制御を回避し、
規約違反のコンテンツを生成する



3 AIサービスの規約内で生成

例：「街中にクマが出没した画像」を生成
※生成後の利用・拡散によっては、誤情報が広がり、
不要な警戒・対応を招く



*1: AP News「Grok blocked from undressing images in places where it's illegal, X says」,https://apnews.com/article/grok-musk-deepfake-nudification-abuse-f0d62ec68576dcfe203cada2424bd107（26/10/16 参照）
*2: 警察庁「令和7年における少年非行及び子供の性被害の状況について」（広報資料）,https://www.npa.go.jp/publications/statistics/safetylife/pdf_r7_syonenkohosiryo.pdf（26/10/16 参照）
*3: 経済産業省「SNS等におけるなりすまし詐欺広告に関する対策強化のための7府省庁合同要請の実施」,https://www.meti.go.jp/press/2026/08/20260807001.html（26/10/16 参照）
*4: WIRED「A Zelensky Deepfake Was Quickly Defeated. The Next One Might Not Be」,https://www.wired.com/story/zelensky-deepfake-facebook-twitter-playbook（26/10/16 参照）
*5: 朝日新聞「AIも自治体も見破れなかった「偽クマ」画像　発端は「内輪の遊び」」,https://www.asahi.com/articles/ASTD310S6TD3UNHB006M.html（26/10/16 参照）
*6: 朝日新聞「全壊の自宅前で笑顔でピース？　被災者の偽写真が拡散、AIで加工か」,https://www.asahi.com/articles/ASV8G3HYRV8GUTIL02TM.html（26/10/16 参照）

Copyright (c) The Japan Research Institute, Limited

日本総研
The Japan Research Institute, Limited

ディープフェイクが金融犯罪にもたらす構造変化

- ディープフェイクは、本人確認突破・口座乗っ取り・投資詐欺・経営者なりすましといった既存の金融犯罪の手段として悪用される
- 攻撃の実行コストを下げる一方、本物らしさを高め、大規模な攻撃を可能にする手段

ディープフェイクを用いた攻撃の広がり

- 詐欺による金銭被害に加え、その手段となる本人確認の突破やなりすましにも、ディープフェイクが用いられた事例が確認されている



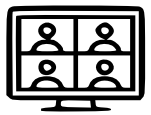
米国のAI関連申告被害額は約8.93億ドル*1

FBIへの2025年の苦情申告ベース
音声クローン、偽ID・偽プロフィール、偽動画などを含む
※AI関連全体の額であり、ディープフェイク単独の被害額ではない



生体認証の不正試行の5件に1件がディープフェイク*2

Entrust社が195か国、10億件超の本人確認を分析
ディープフェイクセルフィー（偽の自撮り）は、2025年に前年比58%増加



国内でオンライン面接におけるAIなりすましが発生*3

実在する経営者の顔と経歴を悪用した応募者が中途採用面接に出現
複数の分析機関がディープフェイクの可能性が高いと判定

ディープフェイクは既存の金融犯罪を巧妙化・大規模化する

ディープフェイクによる能力増幅

- 本人らしい音声・映像の生成
- 偽アカウント・偽の証拠の作成
- 大量・高速・多言語での実行
- 複数モダリティ（テキスト/動画像）を組み合わせ

既存の金融犯罪

経営者・取引先なりすまし	投資詐欺
本人確認・口座開設不正	不正ログイン・不正送金

*1: FBI Internet Crime Complaint Center「2025 IC3 Annual Report」, https://www.ic3.gov/AnnualReport/Reports/2025_IC3Report.pdf（26/10/1 参照）
*2: Entrust「2026 Identity Fraud Report」, https://www.entrust.com/resources/reports/identity-fraud-report（26/10/1 参照）
*3: 読売新聞「IT企業のオンライン採用面接、生成AIで他人になりすまし…実在人物の写真や履歴書悪用か」2026年3月, https://www.yomiuri.co.jp/national/20260319-GYT1T00020/（26/10/1 参照）

Copyright (c) The Japan Research Institute, Limited

日本総研 The Japan Research Institute, Limited

金銭的被害を伴うディープフェイク犯罪の3類型

● ディープフェイク犯罪は、偽られる対象によって「人物・企業」「本人性」「損害・請求証拠」の3類型に整理できる

類型	偽られる対象	主なモダリティ	狙われる場面	被害の発生経路	被害事例
人物・企業のなりすまし	経営幹部、取引先、著名人、政府要人	音声 動画 テキスト	企業の送金承認、SNS広告やWeb会議での投資勧誘	**本人による送金** だまされた本人が自ら振込・入金を実行してしまう	**2024年 香港：**CFOや同僚に見えるWeb会議を信じ、約2,500万米ドル（約40億円）を送金[*1] **2025年 日本：**SNS型投資詐欺（うち著名人をかたる広告経由のものを含む）の年間被害額は約1,288億円[*2]。著名人の偽広告にAI生成の画像・動画が用いられる事例が増加[*3]
本人性の偽装	顔、本人確認書類、契約書類	画像 動画	口座開設、ローン申請、保険契約の手続き	**不正口座・契約の悪用** 他人名義の口座や契約が詐欺やマネロンに使われる	**2024年 インドネシア：**AI加工写真・仮想カメラでデジタルKYCを突破、1社で1,100件超の試行。国内の潜在被害額は約1.4億ドル（約210億円）と推計[*4] **2025/26年 オランダ：**AI加工写真で本人確認を通過し他人名義47口座を開設、銀行損害1.3万ユーロ超（約230万円）[*5]
損害・請求証拠の偽装	事故や損害の画像、請求書、見積書	画像 テキスト	保険金請求、損害賠償請求の審査	**不正請求への支払い** 偽の証拠をもとに保険金・賠償金が支払われる	**2025年 米国 ：**ニューヨークの民泊施設のホストが、加工写真により英国の旅行者に約1.2万ポンド（約240万円）を不当請求[*6] **2026年 英国：** AI画像で架空のキッチン浸水を仕立て、偽造請求書を添えて8万ポンド（約1,700万円）を請求[*7]

*1: World Economic Forum「Cybercrime: Lessons learned from a $25m deepfake attack」, https://www.weforum.org/stories/emerging-technologies/deepfake-ai-cybercrime-arup/（26/10/1 参照）
*2: 警察庁「令和7年における特殊詐欺及びSNS型投資・ロマンス詐欺の認知・検挙状況等について（確定値）」, https://www.npa.go.jp/bureau/criminal/souni/tokusyusagi/sagi_keihatsu2025_teisei.pdf（26/10/1参照）
*3: 内閣官房「SNS等におけるなりすまし詐欺広告に関する対策の強化について（概要）」（資料3-2、なりすまし広告詐欺等対策関係府省庁会議 第1回） 2026年8月, https://www.cas.go.jp/jp/seisaku/narisumashi/dai1/shiryo3-2.pdf（26/10/1参照）
*4: Group-IB「Deepfake Fraud: How AI is Deceiving Biometric Security in Financial Institutions」, https://www.group-ib.com/blog/deepfake-fraud/（26/10/1参照）
*5: Rechtbank Amsterdam「2,5 jaar cel voor oplichten bank met deepfake-technologie」, https://www.rechtspraak.nl/organisatie-en-contact/organisatie/rechtbanken/rechtbank-amsterdam/nieuws/2026/06/25-jaar-cel-voor-oplichten-bank-met-deepfake-technologie（26/10/1参照）
*6: The Guardian「Airbnb guest says images were altered in false £12,000 damage claim」, https://www.theguardian.com/technology/2025/aug/02/airbnb-guest-damage-claim-refund-photos（26/10/1参照）
*7: Allianz Insurance「Posting car pics fuels fraud」 2026年6月5日, https://www.allianz.co.uk/news-and-insight/news/posting-car-photos-fuels-fraud.html（26/10/1参照）

Copyright (c) The Japan Research Institute, Limited

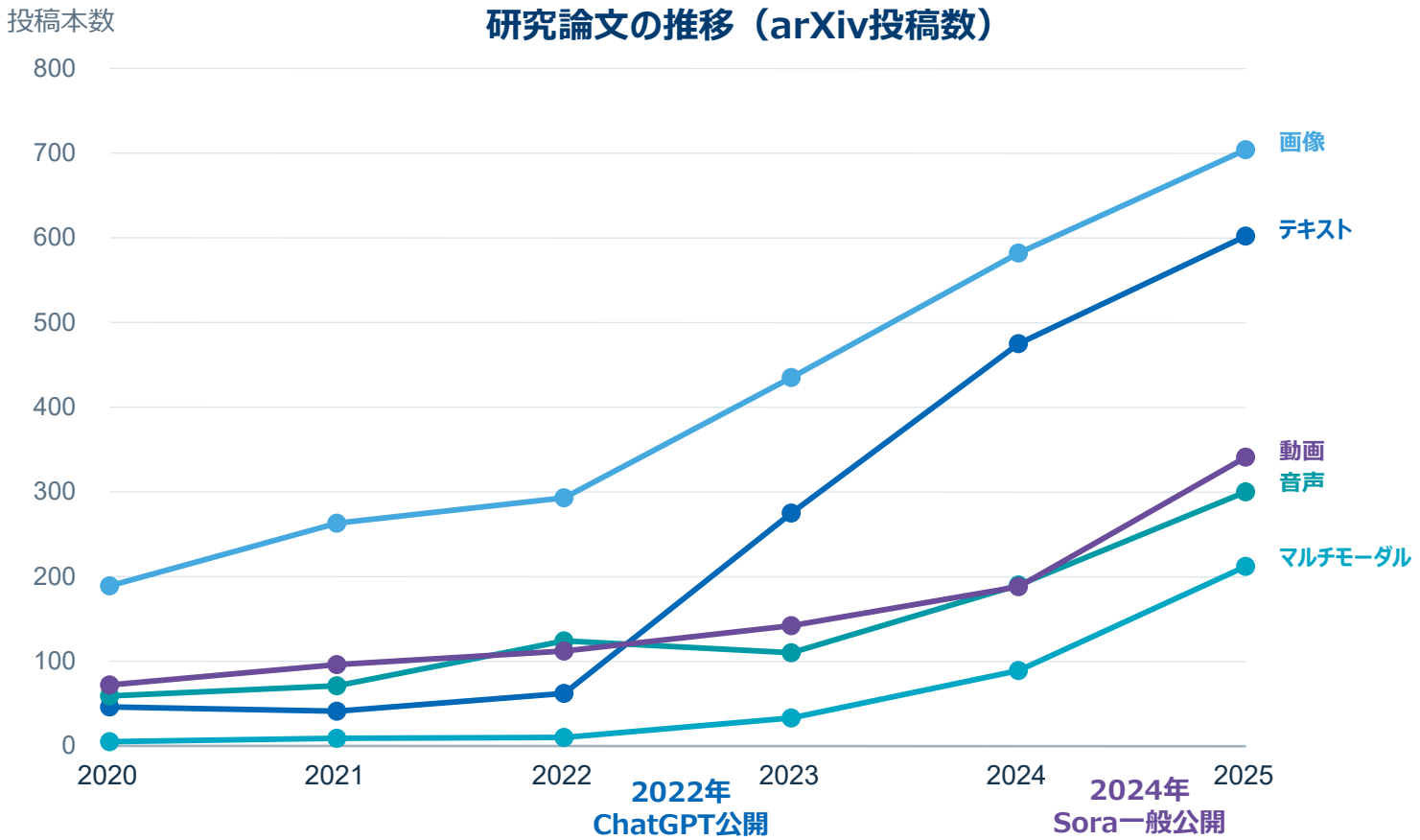
ディープフェイク検出技術の研究動向

- ディープフェイク検出に関する論文数は、テキスト・画像・音声・動画の全モダリティで増加傾向にある
- エポックメイキングな製品（ChatGPT等）の登場が、検出技術の研究を加速させている

モダリティ別の学術研究傾向

テキスト	2022年のChatGPT公開以降に、テキスト関連の研究が急増
画像	2020年時点でも論文数が顕著に多く、現在も堅調に増加している
音声	研究数は継続的に増加している
動画	2024年のOpenAIのSora一般公開以降、研究数が急増している
マルチモーダル	近年急増しており、高性能なVLMの普及が背景にある可能性がある

※分析条件：arXiv APIで、モダリティ別に用意した検索語をタイトルまたは概要に含む論文を検索／対象期間：2020/01/01～2025/12/31（6年間）
arXivメタデータはCC0 1.0
Thank you to arXiv for use of its open access interoperability.



出所：arXiv API（https://info.arxiv.org/help/api/）による検索結果を基にNEC・日本総研にて作図

Copyright (c) The Japan Research Institute, Limited

ディープフェイク検出技術の全体像

- 検出技術は、生成時に検証情報を付与する能動的手法と、生成後にコンテンツを解析する受動的手法に大別できる

能動（プロアクティブ）

生成時に検証可能な情報を付与

生成物にメタデータや電子透かしなどの検証情報をあらかじめ埋め込み、後からその情報を照合して真正性を確かめる



分類（アプローチ）	概要	モダリティ	メリット	デメリット
来歴・署名情報の検証	メタデータに記録された来歴・署名情報を照合して真正性を検証する	テキスト／画像 音声／動画	・生成元や来歴の**明確な証拠として検証しやすい** ・改ざんや加工の有無も検出しやすい	・生成側での情報付与が前提となる ・情報が**除去・欠落すると検証できない**
ウォーターマーク検証	電子透かしとして埋め込まれたウォーターマークを検出・検証する	テキスト／画像 音声／動画		
データベース照合	正規コンテンツを生成・公開時にデータベースへ登録し、後から照合する	テキスト／画像 音声／動画		

受動（リアクティブ）

生成後のコンテンツを解析して検出

出来上がったコンテンツそのものを解析し、AI生成に固有の痕跡や不自然さの有無から真偽を推定する



分類（アプローチ）	概要	モダリティ	メリット	デメリット
AIによるアーティファクト（痕跡）の分析	生成AIに固有のノイズや不自然さを機械学習モデルで検出する	テキスト／画像 音声／動画	・**付与情報がなくても**コンテンツ単体で検証可能 ・生成側の対応を待たず第三者が検証できる	・**判定は確率的で**、誤検出・見逃しが避けられない ・生成技術の進化への継続的な追随が必要
モダリティ間の整合検証	映像と音声など複数モダリティ間の不整合を検証する	動画／ マルチモーダル		
類似検索・出典検証	外部情報との不整合から真偽を推定する	マルチモーダル		

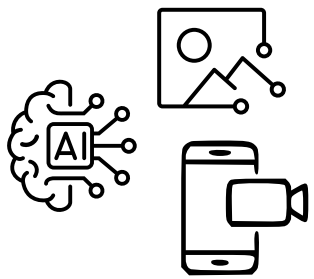
Copyright (c) The Japan Research Institute, Limited

日本総研
The Japan Research Institute, Limited

能動的検出アプローチの基本

- 能動的手法は、生成・編集・撮影の時点で検証に有用な情報をコンテンツに付与し、事後の真正性検証を可能にするもの

1 コンテンツ作成

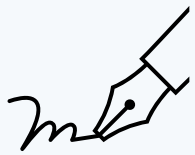


AI生成やカメラ、編集ツールでコンテンツを作成

2 真正性情報を付与・登録

真正性やAI生成の証拠となる情報を付与・登録する

来歴・署名情報の検証



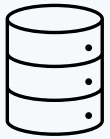
メタデータや電子署名に、著作者・時間情報・編集履歴等を付与し、検証時に情報を確認する

ウォーターマーク検証



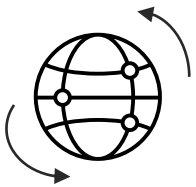
コンテンツ自体に電子的な透かし（ウォーターマーク）を埋め込み、検証時に透かしを確認する

データベース照合



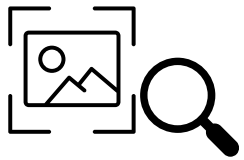
既知のフェイクや正規のコンテンツを登録しておき、検証時に登録済みのデータかを確認する

3 流通



SNSやWeb上でコンテンツが共有・利用される

4 検証・判定



一致：登録情報との整合を確認

不一致・欠落：追加調査や別手段で確認

付与・登録情報を検証し、正規のコンテンツかを確認する

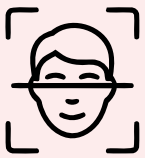
Copyright (c) The Japan Research Institute, Limited

受動的検出アプローチの基本（1/2）：生成AIの痕跡を学習して判定

- 生成AIによるコンテンツは、一見自然であっても、人間には知覚しにくいアーティファクト（痕跡）が含まれる場合がある
- 受動的検出は、この痕跡を大量のデータから学習し、真偽を推定するアプローチである

1 学習段階

大量のフェイクコンテンツと本物のコンテンツを用いて、
アーティファクトのパターンを学習する



フェイクデータ



本物データ

**モダリティや技術方式ごとに、
学習させる痕跡が異なる**

- **テキスト**：語彙・文体 など
- **画像**：画素・階調・色の統計 など
- **音声**：音響特徴・抑揚・リズム など
- **動画**：動き・まばたきの自然さ など

学習済みの検出モデル

2 推論（実用）段階

検証対象のコンテンツ

テキスト　画像

音声　動画

学習済みの
検出モデル

3 出力（判定結果）

- 本物かフェイクかを判定

本物　フェイク

- 方式によっては、疑わしい箇所や
判定根拠も提示

不自然な箇所の
可視化

テキストによる
判定根拠の説明

Copyright (c) The Japan Research Institute, Limited

日本総研
The Japan Research Institute, Limited

受動的検出アプローチの基本（2/2）：コンテンツ内外との整合性検証

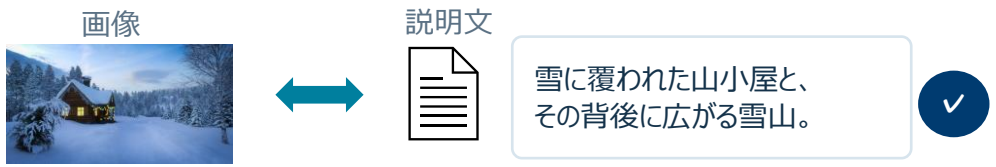
- コンテンツ内のモダリティ間の整合性を確認する方法と、コンテンツ外の証拠（類似物・出典）と照合する方法がある
- いずれも単一の痕跡に依存せず、複数の根拠から真偽を検証する点に特徴がある

モダリティ間の整合検証

画像・音声・テキストの間で、内容や動きが対応しているかを確認する

フェイクコンテンツは、各モダリティ単体では自然でも、組み合わせると矛盾やずれが生じることがある。画像とテキストの両方を扱えるAIである**VLM (Vision-Language Model)**などを用いて、こうした不整合を検出する

例）画像と説明文の対応

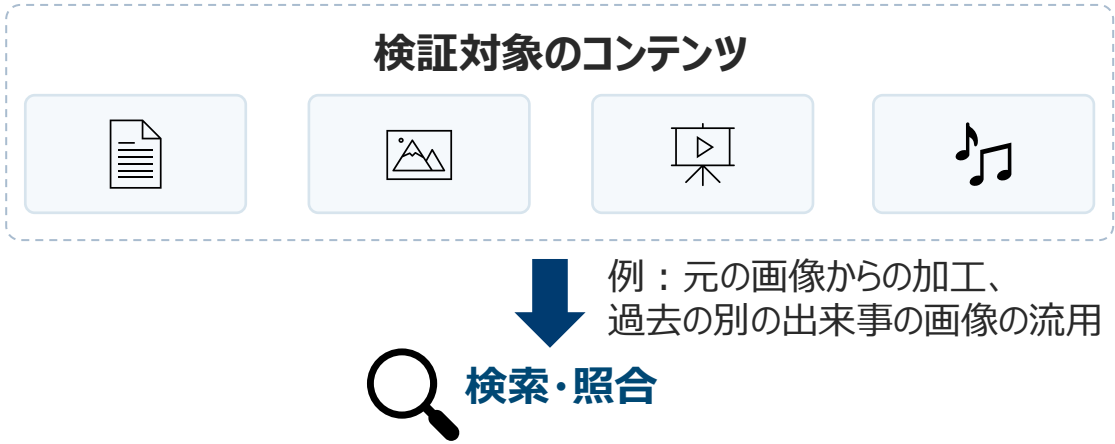


例）映像と音声の対応



類似検索・出典検証

類似コンテンツや元の出典を検索し、内容や掲載文脈に食い違いがないかを確認する



Copyright (c) The Japan Research Institute, Limited

日本総研 The Japan Research Institute, Limited

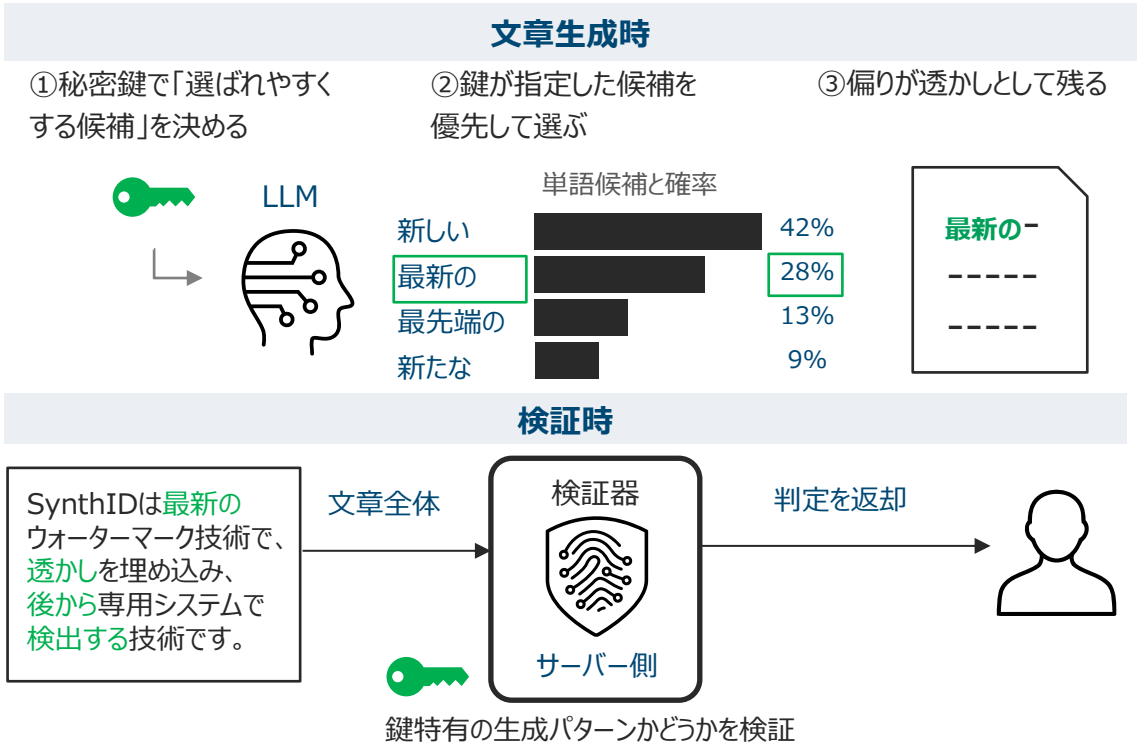
テキストフェイク検出の技術動向

- テキスト領域では、LLMの出力候補に統計的な偏りを埋め込むウォーターマーク技術の実装が進む
- 生成後の判別では、語彙・構造・意味の流れ等の特徴を学習したモデルで真偽を推定する手法が用いられる

能動

統計的ウォーターマーク埋め込みによる生成元の推定

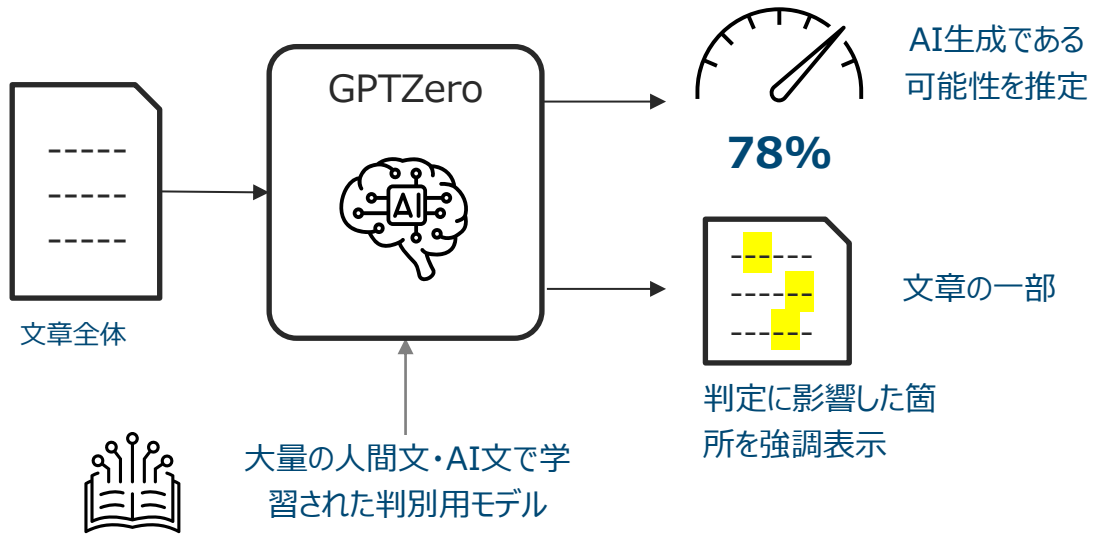
- Google DeepMindが開発したSynthID-Text[*1]が代表例
- GeminiやClaude[*2]も採用し、自社LLMによる生成かを推定できる



受動

学習済み判別モデルによるAI生成テキストの判定

- 語彙・構造・意味の流れなどの特徴を学習したモデルで判定
- GPTZero[*3]は、このような判別モデルを用いたAI生成テキスト検出サービス
- 同社の評価では精度99.76%[*4]。ただし、短文や大幅に書き換えた文章では精度が低下する



*1: Dathathri et al.「Scalable watermarking for identifying large language model outputs」Nature, 2024, https://www.nature.com/articles/s41586-024-08025-4（26/10/1参照）
*2: Anthropic「How Claude's text watermark works」https://www.anthropic.com/news/claude-text-watermark（26/10/26参照）
*3: Adam et al.「GPTZero: Robust Detection of LLM-Generated Texts」arXiv, 2026, https://arxiv.org/abs/2602.13042（26/10/6参照）
*4: Adam・Cui「GPTZero AI Detection Benchmarking」（2026）, https://gptzero.me/news/gptzero-ai-detection-benchmarking-the-industry-standard-in-accuracy-transparency-and-fairness/（26/10/6参照）

Copyright (c) The Japan Research Institute, Limited

画像フェイク検出の技術動向

- 画像領域では、C2PAによる真正性証明を用いた能動的手法の普及が進んでいる
- 受動的手法では、Vision-Language Model（VLM）により判定結果と判断根拠を併せて提示する技術が登場

能動

C2PA技術による真正性証明

- C2PA（Coalition for Content Provenance and Authenticity）
  AdobeやIntel等によるコンテンツ来歴に関する標準化規格・団体
- 本規格は、ソニー製カメラ、Adobe Photoshop、OpenAIの画像生成等に搭載されており、業界標準として普及が進んでいる

C2PAで画像に記録される主な情報（マニフェスト）

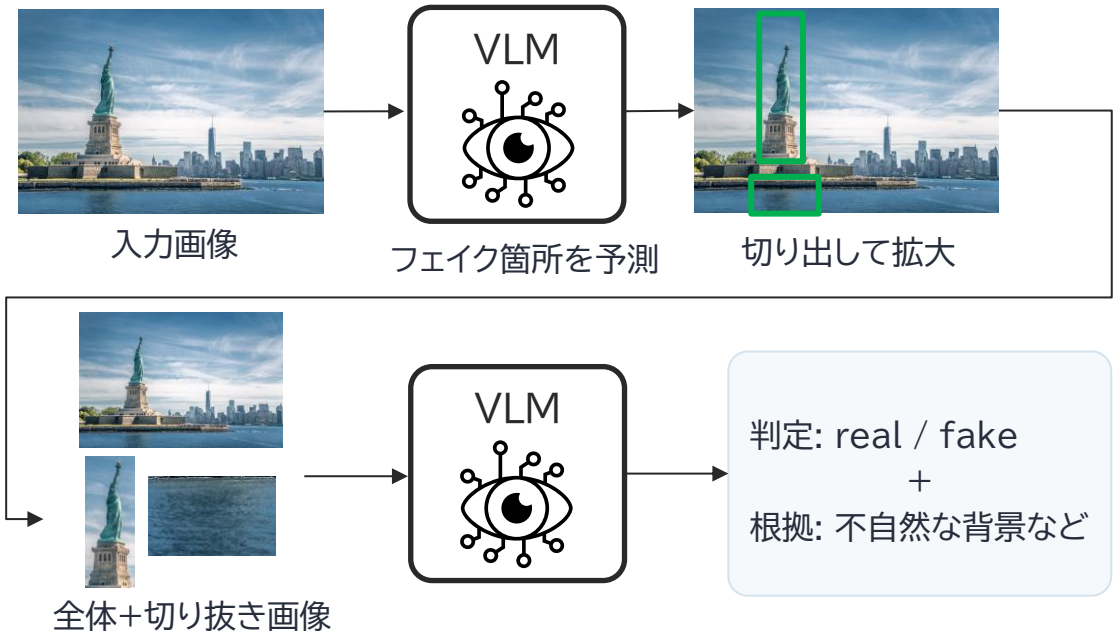
- 撮影日時・位置情報、使用した機器やアプリケーション
- 編集履歴（トリミング、加工、生成AI利用の有無など）
- 発行者の電子署名（改ざんの有無を検証できる）

検証者は、署名の有効性と来歴情報を確認することで、
撮影から流通までの経路が途切れていないかを判断できる
※署名やメタデータが除去・欠落した画像は検証ができなくなる

受動

VLMを利用した画像フェイク検出

- VLM（画像と文章の両方を扱えるAI）を活用。フェイクらしい箇所を切り出して拡大し、全体と局所領域をもとに判定結果と説明を出力する
- 学習に利用していない未知データ（全体生成・部分加工の画像）でも、85～90%程度の検出性能を報告[*1]



*1: Ji et al.「Locate-Then-Examine: Grounded Region Reasoning Improves Detection of AI-Generated Images」CVPR 2026, https://arxiv.org/abs/2510.04225（26/10/1参照）

Copyright (c) The Japan Research Institute, Limited

日本総研 The Japan Research Institute, Limited

音声フェイク検出の技術動向

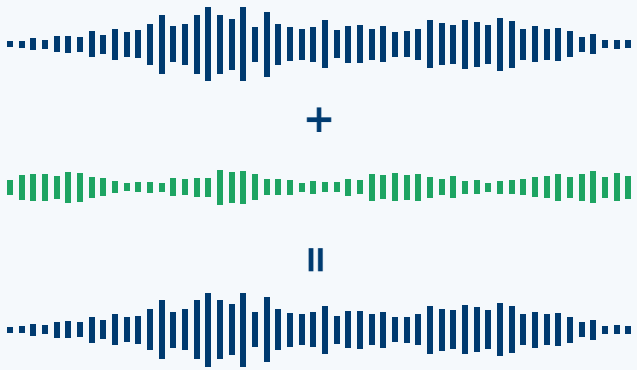
- 音声領域では、波形に人が知覚できない信号を埋め込むウォーターマーク技術の開発が進んでいる
- 受動的手法では、通話中のリアルタイム検出と事後の音声フォレンジックという二つの実装が並行して進展している

能動

波形へのウォーターマーク埋め込みによるAI生成の証明

- AIが生成した音声に、人には聞こえない微細な信号をウォーターマークとして音声波形に直接埋め込む
- 圧縮・切り出し・再エンコードなどの加工を経ても、ウォーターマークが残るよう設計。専用の検出器で、AI生成音声かどうかを事後に検証する
- Google DeepMindが開発したSynthID Audio[*1]は、Google NotebookLM等のポッドキャスト生成などに採用されている

ウォーターマークの埋め込みイメージ



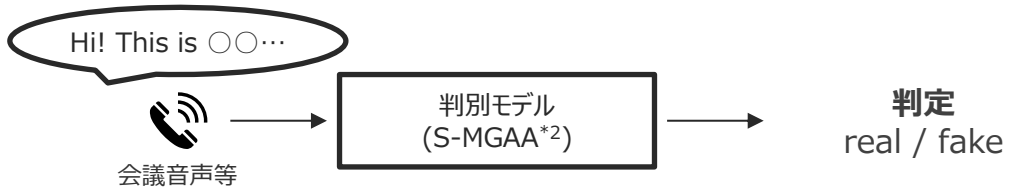
*1: Google DeepMind「SynthID」, https://deepmind.google/models/synthid/（26/10/6参照）
なお、SynthIDはテキスト・画像・音声・動画の透かし技術の総称。音声版がSynthID Audio

受動

受動的検出は、リアルタイム判定と事後フォレンジックの2系統

1 リアルタイム検出（即時判定）

- 英国の大学チームのS-MGAA[*2]は、冒頭の挨拶0.5～2.0秒でフェイクを検出
- 同論文では、軽量な設計と通信劣化への頑健性を報告



適用先：コールセンターや通話窓口など、その場で真偽を判断し警告する場面

2 事後フォレンジック（詳細調査）

- 音声セキュリティ企業Pindrop社の研究[*3]では、どの生成システムで作られたかまで推定
- 入力タイプ99%、音響モデル85～99%、ボコーダー（音声波形の生成）85～97%の識別性能を示す

推定項目	推定結果の例
判定	real / fake
入力タイプ	real / テキスト / 声変換
音響モデル	real / vits / xtts-v1 / …
ボコーダー	real / bark / mul-gan / …

適用先：インシデント調査や監査・鑑定など、事後に詳細な分析が求められる場面

*2: Shi et al.「Audio Deepfake Detection at the First Greeting: “Hi!”」 ICASSP 2026, https://arxiv.org/abs/2601.19573（26/10/6参照）
*3: Klein et al.「Source Tracing of Audio Deepfake Systems」 Interspeech 2024, https://arxiv.org/abs/2407.08016（26/10/6参照）

Copyright (c) The Japan Research Institute, Limited

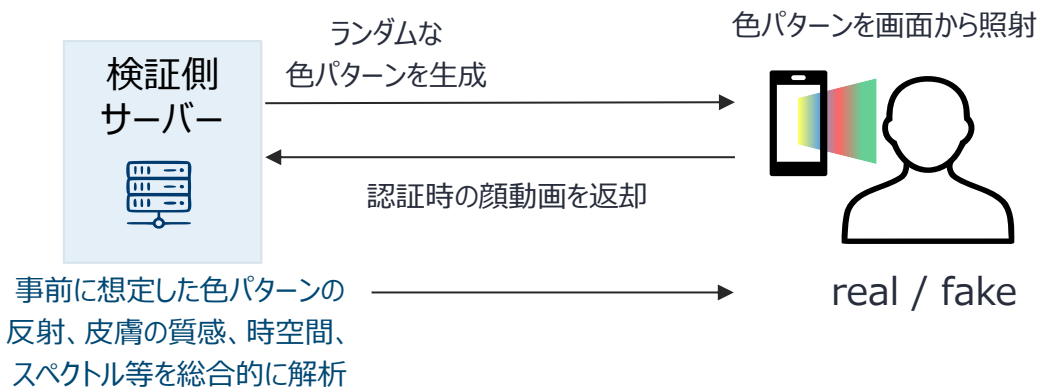
動画フェイク検出の技術動向

- 動画領域では、ディープフェイク対策を組み込んだライブネス検知技術が実用化されている
- 加えて、音声と顔の動きの対応関係を学習した生成AIの内部特徴を、検出に転用する技術も現れている

能動・受動（ハイブリッド※）

色パターン照射によるライブネス検知

- 認証時に予測困難な色パターンを顔に照射し、
  顔の反射と期待する色の系列を照合して、人間であることを確認する技術*1
  ※ 検証側が色パターンを照射して手掛かりを作る「能動」と、
  顔動画に映る光の反射や変化を解析する「受動」を組み合わせたハイブリッドと位置付け
- 代表例：iProov社のFlashmark（同社の認証サービス）
- FIDO認証試験で、本人を誤って拒否する割合は0.14%（同社公表）*2

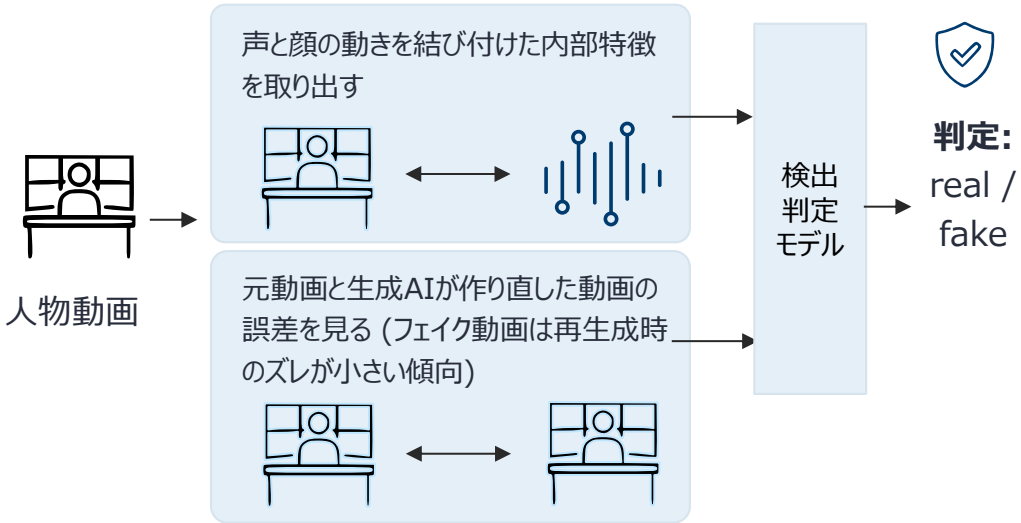


受動

音声と顔の動きの対応関係を活用した動画フェイク検出

- 動画生成AIが学習した「声と口の動きの対応関係」を使って真偽を判定。
  韓国KAISTが開発し、画像認識の国際会議（CVPR 2026）で発表*3。
- 未知の動画生成AIのフェイク動画に対して平均91.98%の検出精度

検査動画を人物動画生成AIに
入力し直し、**内部の反応を解析**



*1: Bud・Newell（iProov）「Biometric methods for online user authentication」US10951613B2, https://patents.google.com/patent/US10951613B2/en（26/10/6参照）
*2:「iProov Becomes First Provider To Achieve FIDO Alliance Certification for Face Biometric Identity Verification」https://www.iproov.com/press/first-provider-fido-alliance-certification-face-biometric-identity-verification（26/10/6参照）

*3: Kim et al.「X-AVDT: Audio-Visual Cross-Attention for Robust Deepfake Detection」CVPR 2026, https://arxiv.org/abs/2603.08483（26/10/6参照）

Copyright (c) The Japan Research Institute, Limited

日本総研
The Japan Research Institute, Limited

マルチモーダルフェイク検出の技術動向

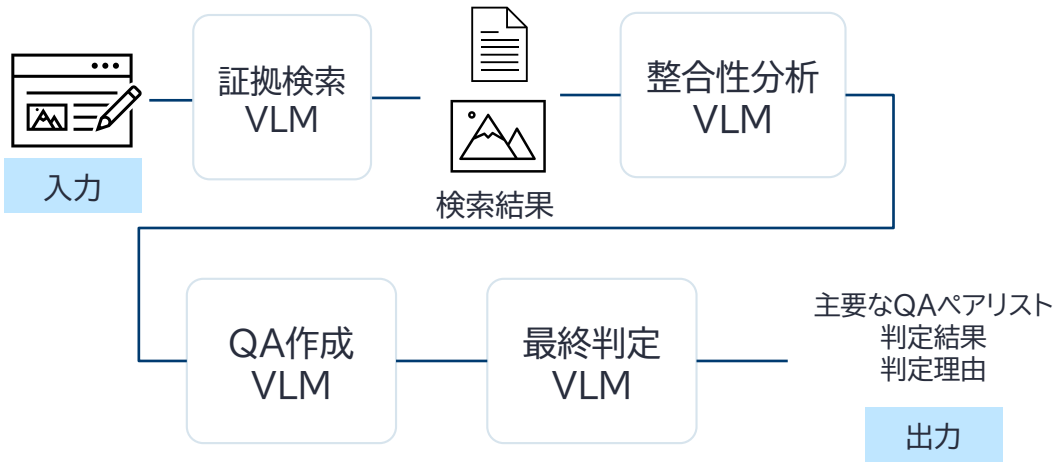
- マルチモーダル領域では、画像とテキストの組み合わせに対し、複数のVLMエージェントが連携して真偽を検証
- 国内でもNECが、任意のモダリティの組み合わせに対応可能な技術を開発している

受動

複数のVLMエージェントの連携によるファクトチェック

- 画像・テキストの組み合わせを、複数のVLMエージェントが分担して検証する技術[*1]
- 証拠の検索・分析、QAペアの作成を経て、真偽を判定する
- 自動ファクトチェックの国際コンペ（FEVER 2026）で1位
  （証拠も含めた厳格な正答率54.6%※）

※判定（支持／否定／証拠不十分／恣意的な情報選択の4分類）と根拠の証拠がともに正しい場合のみ正解とする、コンペ独自の厳しい指標

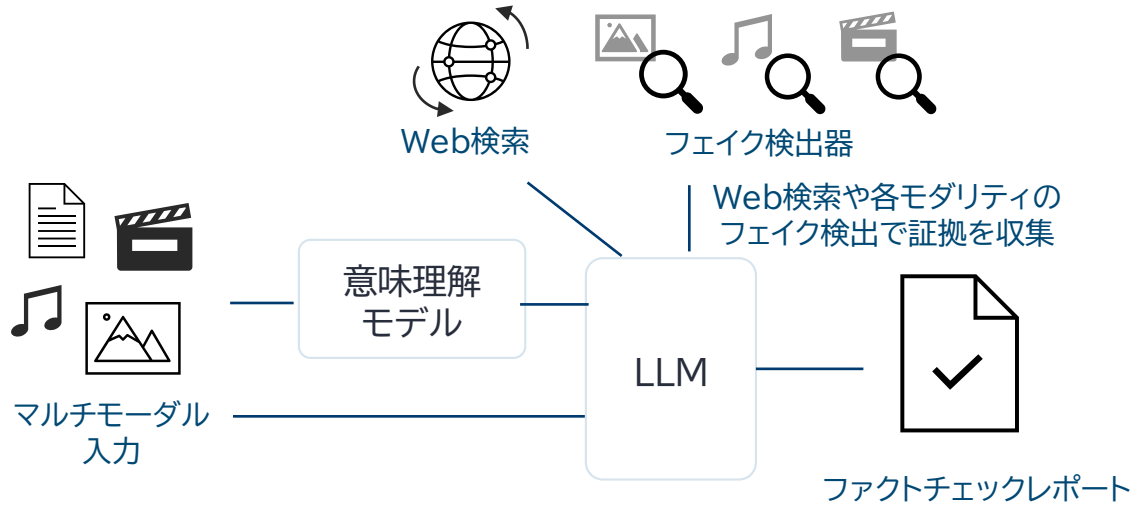


*1: Jung et al.「VILLAIN at AVerImaTeC: Verifying Image–Text Claims via Multi-Agent Collaboration」FEVER 2026, https://aclanthology.org/2026.fever-1.9/（26/10/7参照）

受動

マルチモーダルな情報をテキスト化しLLMで処理

- 画像・映像・音声・テキストの任意の組み合わせに対し、LLMでファクトチェックを行う技術[*2]。AI分野の国際会議（AAAI 2025）で発表
- NECが開発し、実用化に向け実証中
- 各モダリティをテキスト化してLLMで処理し、任意の組み合わせに対応
- 改良版（NEC・東京大学）は実世界のデータで正答率78%[*3]



*2: Kakizaki et al.（NEC）「MAFT: Multimodal Automated Fact-Checking via Textualization」AAAI 2025, https://ojs.aaai.org/index.php/AAAI/article/view/35354（26/10/7参照）
*3: 松永ほか（NEC・東京大学）「MAFT-AH: ヒューマンインザループを備えたマルチエージェント自動ファクトチェックフレームワーク」言語処理学会 第32回年次大会, https://www.anlp.jp/proceedings/annual_meeting/2026/pdf_dir/P2-6.pdf（26/10/7参照）

Copyright (c) The Japan Research Institute, Limited

日本総研 The Japan Research Institute, Limited

先端的な検出手法のトレンドまとめ

- 検出技術は「真正性の活用」・「説明可能性」・「汎化」・「マルチモーダル活用」の4つの方向で進展している
- 単独の検出器による二値判定から、複数の証拠に基づき真正性と判断根拠を示す検証へと移行しつつある*1, *2

真正性の活用

従来：事後検出のみ

▶ 生成時・撮影時に付与した真正性情報も活用

後から検証するだけでなく、正規の証拠を事前に付与

- テキスト・音声へのウォーターマーク埋め込み（P.12、P.14）
- 署名付きコンテンツ来歴規格C2PAの活用（P.13）
- 色パターン照射によるライブネス検知（P.15）

説明可能性

従来：リアル/フェイクの二値判定

▶ 判断根拠や由来も推定

フェイクと出力するだけでなく、判断根拠・フォレンジックも提示

- 判定に影響した箇所を強調表示（P.12）
- VLMによる説明可能な画像フェイク検出（P.13）
- 音声の生成元推定（P.14）

汎化

従来：既知データでの高精度化

▶ 未知の生成器・実環境への汎化

未知の生成器・実環境・リアルタイム性を考慮して検出

- 通話環境を想定した早期フェイク音声検出（P.14）
- 音声と顔の対応関係に着目し、未知の生成AIにも対応（P.15）

マルチモーダル活用

従来：単一モダリティのみの分析

▶ マルチモーダル・外部証拠の活用

複数モダリティや外部証拠を横断して整合性を確認

- 証拠検索に加え、複数VLMが連携して検証（P.16）
- 証拠検索に加え、マルチモーダル情報をテキスト化して検証（P.16）

*1: Deng et al.「A Survey of Defenses Against AI-Generated Visual Media: Detection, Disruption, and Authentication」ACM Computing Surveys, 2025, https://arxiv.org/abs/2407.10575（26/10/7参照）
*2: Lin et al.「Detecting Multimedia Generated by Large AI Models: A Survey」arXiv, 2024, https://arxiv.org/abs/2402.00045（26/10/7参照）

Copyright (c) The Japan Research Institute, Limited

日本総研 The Japan Research Institute, Limited

国内外における検出技術の社会実装状況

- 海外では、検出技術が本人確認・コールセンター・オンライン会議などの業務プロセスに組み込まれ、政府調達も始まっている
- 国内は、総務省とNEDOの公的事業のもとで、インターネット上の偽・誤情報を対象とした技術の開発・実証が進む

海外：金融・政府の業務プロセスにも組み込まれ始めている

- 当局の注意喚起*1を受け、欧米の金融・政府機関で製品導入が先行

	Hive（米国）	iProov（英国）	Pindrop（米国）
判定内容	AI生成コンテンツ	実在する本人	合成音声・なりすまし
対象	画像・動画・音声・テキスト等のコンテンツ全般	顔画像・ライブ映像	通話の音声・会議映像
扱う業務	AI生成コンテンツ・顔ディープフェイクの判定、捜査での優先順位づけ	オンライン本人確認・口座開設、顧客認証、アカウント復旧時のなりすまし防止	コールセンターでの詐欺電話の防止、音声の事後鑑定、会議の偽参加者検出
導入・提携の例*2	・米国防総省DIUが採用（24年12月） ・米国土安全保障省が捜査で試験導入（25年9月）	・スイスUBSが口座開設に導入（22年11月） ・比UnionDigital Bankが本人認証に導入（25年10月）	・米銀FNBOが通話の不正検出に活用 ・Zoom等で一般提供（25年11月）

国内：公的事業のもとでの開発・実証の段階

- オンライン本人確認は、制度での対策が進む
  - 口座開設等は2027年4月に画像送信型を廃止、マイナンバーカード等のICチップ読み取りを中心とする方式へ移行*3。**金融庁は、施行日を待たず対応を進めるよう金融機関等に要請*6、業界では画像送信型の早期廃止が進む**
- 検出技術は、**SNS等の偽情報や通話のなりすまし対策**で開発・実証

取り組みを支える公的事業
- 総務省「偽・誤情報対策技術の開発・実証事業」（2026年度、13者採択）*4
- NEDO「K Program」偽情報分析技術の開発（2024～27年、9者参画）*5

採択・参画した事業者の例

	NEC	富士通	NABLAS
開発・実証内容	AIを活用したマルチモーダルな偽・誤情報の分析技術	偽情報の検知から根拠収集、分析、評価までの統合基盤の構築	AI合成音声等による通話なりすましの検出
想定する使い手	ファクトチェック機関、自治体、マスメディア等	公的機関、民間企業等	通信事業者、金融機関、警察等
公的事業	総務省事業	NEDO K Program	総務省事業

*1: FinCEN「FinCEN Alert on Fraud Schemes Involving Deepfake Media Targeting Financial Institutions」2024, https://www.fincen.gov/sites/default/files/shared/FinCEN-Alert-DeepFakes-Alert508FINAL.pdf（26/10/7参照）
*2: 各社・関係機関の公表資料、報道をもとに作成
*3: 警察庁「犯罪による収益の移転防止に関する法律施行規則の一部を改正する命令」（令和7年6月24日公布）, https://www.npa.go.jp/sosikihanzai/jafic/hourei/250624/anbun02.pdf（26/10/20参照）
*4: 総務省「『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』の公募結果」2026, https://www.soumu.go.jp/menu_news/s-news/01ryutsu02_02000479.html（26/10/7参照）
*5: NEDO「『経済安全保障重要技術育成プログラム』で偽情報の分析に係る技術の開発に着手」2024, https://www.nedo.go.jp/news/press/AA5_101763.html、NII 2024, https://www.nii.ac.jp/news/release/2024/1016.html（26/10/7参照）
*6: 金融庁「現下の情勢を踏まえたサイバーセキュリティ対策の強化と各種取引申込等における対応について」2026, https://www.fsa.go.jp/news/r8/sonota/20261009/20261009.html（26/11/5参照）

Copyright (c) The Japan Research Institute, Limited

日本総研 The Japan Research Institute, Limited

金融詐欺に対する各国のディープフェイク規制

- 詐欺の初期段階（接触・勧誘）はディープフェイクの影響を受けやすい。各国は、生成AIサービス提供者への規制を強化。
- 台湾は広告・金融等の事業者も先行して規制し、日本でも台湾型の規制を目指す提言が発出*1

ディープフェイク規制による対応範囲

既存規制の対応範囲

①AIによるコンテンツ生成	②偽広告・偽アカウントの作成	電話・SMS・ビデオ通話による接触	認証情報の窃取・アカウント乗っ取り	新規送金先の登録・送金	資金回収
• AIコンテンツへのラベル付与 • メタデータの付与 • ウォーターマークによる来歴証明の明示	• 広告主の確認 • ディープフェイク広告の迅速な削除	• AI音声利用の開示 • 正規番号の認証 • 発信番号の偽装防止 • 既知の詐欺番号の遮断	• フィッシング耐性の高いMFAの導入 • 連絡先や認証設定の変更監視・リアルタイム通知 • 利用者による口座取引停止	• 口座名義の照合による警告 • 新規送金先の通知 • 行動分析による異常検知 • 高リスク取引の警告・制御	• 詐欺口座の情報共有 • 口座凍結 • 取引の取り消し

規制	国・地域	施行	概要	事業者区分
EU AI Act	EU	2026.8	機械可読な表示の提供、ディープフェイクであることの表示などを義務化	生成・提供（①）
カリフォルニア州 AI透明性法	米国	2026.8	AI生成検出ツールの提供、来歴情報の付与などを義務化	生成・提供（①）
IT Amendment Rules	インド	2026.2	合法なAI生成情報のラベル・来歴情報の提供、違法なAI生成情報の生成・公開・送信の防止などの義務化	生成・提供（①） SNSなどの投稿・配信サービス
AI基本法	台湾	2026.1	AIの開発・提供・利用に伴う安全性、透明性、データ保護など	（横断的な基本原則）
ペンシルベニア州法	米国	2025.9	ディープフェイク作成の犯罪化、特に金銭詐取については重罪化	（偽造・配布を目的とした利用者）
詐欺犯罪危害防制条例	台湾	2024.8	広告主・資金提供者の本人確認、ディープフェイク広告の表示、24時間以内の詐欺広告削除・アカウント停止、金融口座の詐欺悪用防止などの義務化	広告・金融（②）
生成型AIサービス管理暫定弁法	中国	2023.8	適法な学習データやモデルの利用、個人情報の保護、生成物の表示などの義務化	生成・提供（①）
深層合成管理規定	中国	2023.1	AI生成コンテンツの表示義務	生成・提供（①）

*1: 自民党「「対応しない事業者を放置しない」ディープフェイク対策PTが提言申し入れ」2026, https://www.jimin.jp/news/policy/213332.html（26/9/15参照）

Copyright (c) The Japan Research Institute, Limited

日本総研
The Japan Research Institute, Limited

生成AI大手による透かしの付与と判定ツールの公開

- 2025～26年に、EU・米カリフォルニア州では新たな表示義務が、中国では既存の表示義務を具体化する規則が相次いで施行
  これを受け、 大手AIサービス提供者は生成物への透かし・来歴情報（C2PA）の付与を標準仕様とした
- 一方、透かしの除去・回避への耐性や、対応していない事業者が残る点に課題

規制	施行	生成AI事業者への主な義務	事業者の対応例
EU AI Act 第50条	2026年8月 （既存システムは12月まで猶予）	生成物に、AI生成・加工であることを示す情報を付与	OpenAIは、画像にC2PAとSynthID、音声にもSynthIDの透かしを付与。EU域内のテキストには、自社方式の透かしを導入[*1]
米国 カリフォルニア州 AI透明性法	2026年8月	月間利用者が100万人を超える事業者に、生成物へのAI生成・加工情報の付与と、無料の判定ツールの提供を義務化	OpenAIが判定ツールを公開・API提供[*1]。GoogleはOpenAI等の透かしも判定できる「SynthID Detector」を一般公開[*2]
中国 AI生成合成コンテンツ標識弁法 ※P19に記載の中国国内の規則を具体化	2025年9月	生成物への明示的な表示と、メタデータへのAI生成情報の記録	DeepSeekが生成物への表示付与を告知。WeChatなどもAI生成物への表示や標識の読み取りを導入[*3]

*1: OpenAI「より安全で透明性の高い AI エコシステムに向けて、コンテンツ来歴の取り組みを前進」2026, https://openai.com/ja-JP/index/advancing-content-provenance/（26/10/9参照）
*2: Google「Google expands SynthID Detector for AI content」（2026年10月7日）, https://blog.google/innovation-and-ai/models-and-research/google-deepmind/synth-id-ai-content/（26/10/20参照）
*3: 新華網「《人工智能生成合成内容标识办法》正式施行 多平台出台内容管理细则」（人民郵電報より転載）2025, https://www.news.cn/tech/20250909/fb164c6d092146aa8e13ddc283fe416a/c.html（26/10/9参照）

Copyright (c) The Japan Research Institute, Limited

脅威の進化に対する検出技術の課題と対応の方向性

- 業務への組み込みや規制対応が進む一方、生成AIの進化や加工の巧妙化に検出技術の対応が遅れる
- 特に、汎化性能・実環境への適応・回避攻撃への耐性・説明の信頼性の4点に課題が残る

現状の課題	求められる対応
新しい生成AIが出るたびに、検出が後手に回る • 未知の生成モデルによるディープフェイクでは、検出性能が依然として低下する*1	**継続的に更新・再評価可能な検出基盤の構築** • 自己教師あり学習・継続学習により新しい生成器へ適応 • 新モデルへの更新後も、過去の検出能力を維持できる仕組みを構築*2
実際の業務環境では、論文で報告された性能を下回る • スマートフォンのAI補正などで、「リアル画像」自体が変化している*3 • SNS等で圧縮・加工されて出回った偽コンテンツでは、性能が大きく低下*4	**実環境を想定した検出器の検証** • 新しい生成器・AI補正・SNS加工を継続的に追加するベンチマークの整備 • リアルタイム対応・エッジ端末での実装を見据えた軽量化
検出回避を目的とした加工や透かしの除去への耐性が低い • 高度な加工（敵対的攻撃）により、検出を回避される可能性がある • 規制が前提となるウォーターマークも、編集や攻撃により除去されるおそれ*5	**頑健性を高めた多層検証** • ウォーターマーク・C2PA・受動的検出器の併用 • 必要に応じて、ブロックチェーンなどの改ざん耐性の高い記録基盤を併用*1
判定理由の説明そのものが誤り得る • 自然言語の説明が、実際の判定根拠を反映しているとは限らない • 説明自体にハルシネーションが生じる可能性がある	**説明の信頼性を検証・評価できる仕組み** • 判定の確信度が低い場合、その不確実性を利用者に提示する*1 • 信頼できる外部情報と照合し、判定根拠を裏付ける

*1: Croitoru et al.「Deepfake Media Generation and Detection in the Generative AI Era: A Survey and Outlook」ACM Computing Surveys, 2026, https://arxiv.org/abs/2411.19537（26/10/20参照）
*2: Fontana et al.「Revisiting Deepfake Detection: Chronological Continual Learning and the Limits of Generalization」arXiv, 2025, https://arxiv.org/abs/2509.07993（26/10/7参照）
*3: Keuper & Keuper「Position: We need to re-think the concept of “real” images」ICML 2026, https://arxiv.org/abs/2509.21864（26/10/7参照）
*4: Chandra et al.「Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024」CVPR Workshops, 2026, https://arxiv.org/abs/2503.02857（26/10/7参照）
*5: Zhao et al.「SoK: Watermarking for AI-Generated Content」IEEE S&P, 2025, https://arxiv.org/abs/2411.18479（26/10/7参照）

Copyright (c) The Japan Research Institute, Limited

日本総研 The Japan Research Institute, Limited

金融取引の防御に向けた多層的な対策

- 透かし・来歴の導入は進むが検出技術の性能には課題が残り、生成AIの高度化で単独での防御は一層困難に
- 検出技術はリスクシグナルの一つとして、既存の認証・不正検知や業務手順と組み合わせた多層防御に組み込むべき

取り得る対応策

➢ **能動・受動とも単独の検出技術では限界があり、場面に応じて組み合わせる**

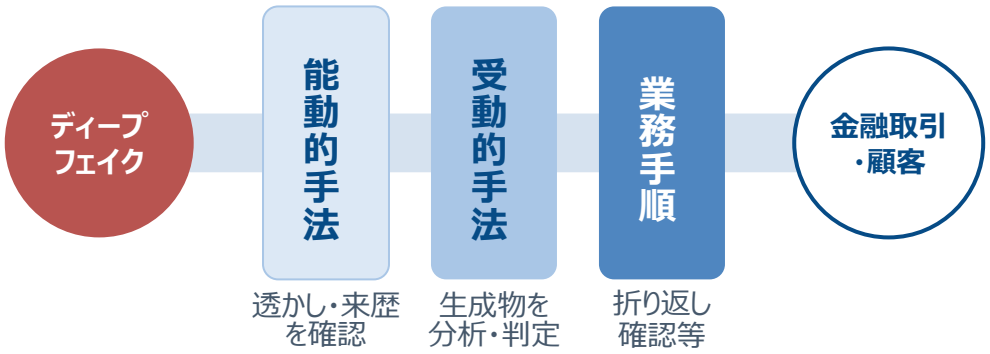
- 能動的手法（透かし・来歴）は導入が進むが、除去や非対応モデルで無効化されうる
- 受動的手法は、透かしなしでも使えるが、未知のモデルや実環境で精度が低下

➢ **技術的手法に加え、業務手順による確認が被害防止の要となる**

ユースケース	対策の組み合わせ例
高額送金・法人取引	ビデオ会議・通話のフェイク検出＋別経路確認＋複数人承認
オンライン本人確認	ICチップ読み取り＋ライブネス検知＋撮影端末の真正性確認
コールセンター	音声フェイク検出＋登録番号への折り返し

➢ **中長期的には真正性証明の普及により、「偽物の検出」から「本物の確認」へ**

- コンテンツに真正性証明（C2PA等）を付与し、社会的に「本物の確認」が優先されることも重要



制度的基盤（表示義務、実環境下での性能評価等）

▲ **行政・公的機関／金融機関・企業／技術開発者による継続的な整備・更新**

提言

- **行政・公的機関**は、表示義務の導入・拡大に加え、研究機関・業界と連携し、実環境を反映したベンチマーク等の共通評価基盤を整備する※
- **金融機関・企業**は、公表精度に依拠せず自社の実環境で検証したうえで、不正検知や業務手順への組み込みを前提に導入を判断する
- **技術開発者（ベンダー等）**は、新たな生成AIや検出回避手法に対応するため、継続的な更新を前提とした開発・運用体制を構築する

※ 先行例として、米国NISTがディープフェイク検知の評価プログラム「Guardians of Forensic Evidence」を推進しており、米国のAI行動計画（America's AI Action Plan）は、同プログラムを正式なガイドライン・任意のベンチマークへ発展させることの検討を求めている。

Copyright (c) The Japan Research Institute, Limited

おわりに

本レポートでは、ディープフェイクがもたらす金銭的被害の実態と、検出技術の全体像と最新動向、社会実装や各国規制の状況、今後の課題を整理した。検出技術は能動的・受動的の両アプローチで着実に進展している一方、単独で被害を防ぎきることは難しい。対策としては、複数の検出技術の併用に加え、本人確認や取引承認などの業務手順、制度的な裏付けを組み合わせた多層防御が重要になる。

生成AIと検出回避手法は今後も高度化し、検出技術との攻防は継続すると見込まれる。そのため、一度講じた対策も固定的なものとせず、新たな手口や技術・規制の動向に応じて随時見直していく必要がある。

ディープフェイクへの対策には、技術ベンダー、金融機関、規制当局の連携が欠かせない。本レポートが、関係者による取り組みの検討の一助となれば幸いである。

Copyright (c) The Japan Research Institute, Limited

技術調査協力

本レポートの執筆にあたっては、ディープフェイク検知や偽・誤情報分析に豊富な知見を有する
日本電気株式会社 セキュアシステムプラットフォーム研究所　AIトラスト研究チームより、多大なご協力をいただきました

古川　諒 日本電気株式会社 リードリサーチエンジニア　修士（工学）

2008年東京工業大学大学院知能システム科学専攻修士課程修了。同年日本電気株式会社入社。
現在AIセキュリティや偽・誤情報対策に関する研究開発に従事。2006年日本機械学会畠山賞受賞。

柿崎　和也 日本電気株式会社 主任研究員　博士（工学）

2017年筑波大学大学院システム情報工学研究科修士課程修了。同年日本電気株式会社入社。
2025年同大学大学院システム情報工学研究群博士後期課程修了。現在AIセキュリティや偽・誤情報対策に関する研究開発に従事。
MITテクノロジーレビュー「Innovators Under 35 Japan 2025」に選出。

松永　悠斗 日本電気株式会社 主任　博士（情報科学）

2023年北海道大学大学院情報科学院情報科学専攻博士後期課程修了。同年日本電気株式会社入社。
現在マルチモーダルフェイク検知の研究に従事。2020-2023年学術振興会特別研究員を歴任。
2023年北海道大学大学院学院長賞受賞。電子情報通信学会高機能マルチメディア研究専門委員会幹事補佐。

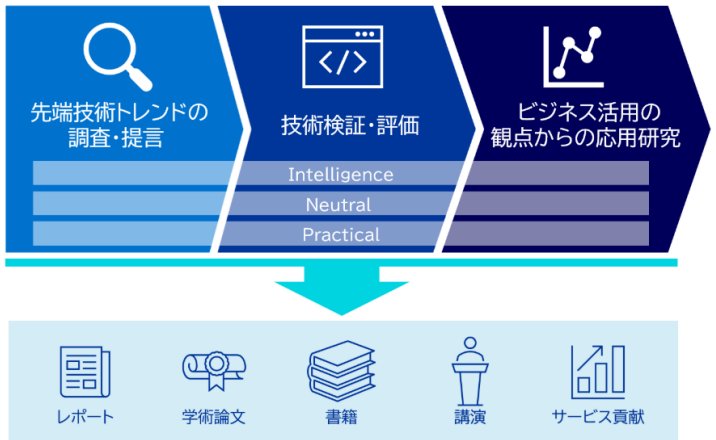
大磯　秀幸 日本電気株式会社 研究員　修士（工学）

2025年筑波大学大学院 理工情報生命学術院システム情報工学研究群情報理工学位プログラム修了。同年日本電気株式会社入社。
現在AIセキュリティおよびディープフェイク検知に関する研究に従事。コンピュータセキュリティシンポジウム2023学生論文賞受賞。

Copyright (c) The Japan Research Institute, Limited

先端技術ラボ ご紹介

先端技術を活用したITサービスの創出に向けた技術の目利き役として、「先端技術トレンドの調査・提言」、「技術検証・評価」、「ビジネス活用の観点からの応用研究」に取り組んでいます



主な研究・取り組み領域



当社ホームページの 特集サイト では、IT分野における先端技術の調査レポート、及び所属する部員のプロフィール詳細がご覧いただけますので、ぜひご参照ください。

本レポート執筆者へのメディア取材や講演などに関するご相談につきましては、当社ホームページの お問い合わせフォーム よりご連絡ください。

株式会社日本総合研究所

日本総研は、シンクタンク・コンサルティング・ITソリューションの3つの機能を有するSMBCグループの総合情報サービス企業です。

東京本社　〒141-0022　東京都品川区東五反田2丁目18番1号　大崎フォレストビルディング

大阪本社　〒550-0001　大阪市西区土佐堀2丁目2番4号

Copyright (c) The Japan Research Institute, Limited