

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

# SNSユーザー支援を中核とした偽・誤情報対策技術の開発・実証

## 成果報告書

2026/3/19

技08\_株式会社データグリッド

# 目次

1. 開発・実証のサマリ
  1. 開発・実証のサマリ
2. 開発・実証の背景・目的
  1. 開発技術によりアプローチする課題
  2. 開発技術により目指す姿・ゴール
  3. 開発技術により対処可能なユースケース
3. 開発・実証における「対策技術の開発」
  1. 技術開発の全体像
  2. 技術開発の個別詳細
4. 開発・実証における「対策技術の有効性等に関する検証及び調査」
  1. 検証及び調査の全体像
  2. 検証及び調査の個別詳細
5. 開発・実証における「対策技術の社会実装に向けた取組」
  1. 社会実装に向けた取組の全体像
  2. 社会実装に向けた取組の個別詳細
6. 開発・実証における「普及啓発活動への協力」
  1. 普及啓発活動の全体像
  2. 普及啓発活動の個別詳細
7. 開発・実証の課題・展望
  1. 技術開発及び社会実装における課題・展望
8. 開発・実証の実施体制等
  1. 実施体制及び役割分担
  2. 全体スケジュール

# 目次

1. 開発・実証のサマリ
  1. 開発・実証のサマリ
2. 開発・実証の背景・目的
  1. 開発技術によりアプローチする課題
  2. 開発技術により目指す姿・ゴール
  3. 開発技術により対処可能なユースケース
3. 開発・実証における「対策技術の開発」
  1. 技術開発の全体像
  2. 技術開発の個別詳細
4. 開発・実証における「対策技術の有効性等に関する検証及び調査」
  1. 検証及び調査の全体像
  2. 検証及び調査の個別詳細
5. 開発・実証における「対策技術の社会実装に向けた取組」
  1. 社会実装に向けた取組の全体像
  2. 社会実装に向けた取組の個別詳細
6. 開発・実証における「普及啓発活動への協力」
  1. 普及啓発活動の全体像
  2. 普及啓発活動の個別詳細
7. 開発・実証の課題・展望
  1. 技術開発及び社会実装における課題・展望
8. 開発・実証の実施体制等
  1. 実施体制及び役割分担
  2. 全体スケジュール

# 1-1. 開発・実証のサマリ

<p>アプローチする課題・目指す姿</p>	<p>SNS上では偽・誤情報が巧妙に投稿・拡散される一方、ユーザーが自ら検証する習慣・ツールが不足し、ディープフェイクの急速な進化や訂正情報の流通不足が深刻化している。本事業では、SNSユーザーの主体的な情報検証を支援するファクトチェック支援アプリ「シラベル」と訂正主体向けSNSモニタリングツールを開発・実証し、社会全体で偽・誤情報に対処できる健全な情報流通環境の構築を目指す。</p>		
<p>技術区分</p>	<p>コンテンツの真偽判別支援技術</p>	<p>実施体制 (下線: 技術開発主体)</p>	<p>株式会社データグリッド、株式会社AIBOS、AIBOS Uganda Co., Ltd.</p>
<p>対象とするモデル種</p>	<p>文章、画像、音声、動画</p>		

## 技術開発の取組・成果

- **技術1 ディープフェイク検出:** 最新の生成AIにより作成された偽の画像・動画・音声を自動判定し、検出精度 93% を達成 (目標90%)
- **技術2 根拠情報の自動検索:** SNS上の真偽不明な言説に対し信頼性の高い複数の情報源から根拠を自動提示し、適合率 92% を達成 (目標85%)
- **技術3 画像・動画の出所検証:** 類似画像・動画検索により初出元や過去の利用文脈をWeb上から特定・提示し、適合率 90% を達成 (目標80%)
- **技術4 情報摂取行動の可視化:** SNS上でのユーザーの情報閲覧傾向や偽情報への反応・拡散の履歴を分析して提示し、精度(QWK) 0.62 を達成
- **技術5 リスク評価・拡散予測:** SNS上の真偽不明言説の社会的影響度を定量的に評価・対応優先順位付けを支援し、精度(QWK) 0.84 を達成

## 実証の取組・成果

- **検証1 SNSユーザー向けアルファ版テスト:** 59名のモニターにアプリを試用いただき、ファクトチェック支援の有用性とリテラシー向上効果を検証。満足度は10段階中 7.81 と高評価を獲得し、利用前後の情報リテラシーテストでは平均 +3.3点 の統計的に有意な向上を確認。
- **検証2 SNSユーザー向けベータ版テスト:** 上記アプリの改善版を一般公開し、普及ポテンシャルとユーザー評価を検証。ポジティブ評価が 76.3% と高い支持を得ており、リリース後1ヶ月で約200名のユーザーを獲得。ファクトチェックニーズを確認するとともに、認知度向上に向けたメディアでのPRが今後重要となると認識。
- **検証3 訂正主体向けテスト:** 報道機関にSNSモニタリングツールをテスト利用いただき、実務への適合性を検証。ユーザービリティ評価は10段階中 7 を獲得し、早期探知・拡散分析・取材効率化など実践的なユースケースを検証。

## 技術開発及び社会実装にあたっての課題・展望

- 技術開発面では、①生成AI技術の急速な進化への追従 (本年度構築したMLOpsパイプラインの自動化率をさらに向上させ、新種ディープフェイクへの対応リードタイムを短縮)、②分析コストの最適化 (軽量モデルの活用やキャッシュ機構の導入)、③検出対象メディアの拡大 (ショート動画やライブ配信への対応)、④多言語対応の高度化 (越境的な偽情報キャンペーンに対応するための多言語根拠検索の精度向上) が主要課題として明確化された。
- 社会実装面では、①SNSユーザーの獲得・定着 (ゲーミフィケーションなどにより日常の行動フローに溶け込むUX設計を実現)、②訂正情報の流通促進 (「シラベル」を訂正情報の流通チャネルとして活用し、訂正情報の到達率を向上)、③訂正主体のツール活用促進 (パートナー企業との協業による既存顧客基盤を活用した間接的な導入促進)、④法規制・倫理面への対応 (あくまで「検証支援」として最終判断はユーザーに委ねる設計思想の堅持) に取り組む。

## 代表者コメント



株式会社データグリッド  
代表取締役  
岡田 侑貴

SNS上の偽・誤情報に対抗する最大の鍵は、私たち一人一人の情報リテラシーにあると考えています。生成AIで巧妙な情報が溢れる時代だからこそ、情報の真偽を丁寧に見極める慎重かつ冷静な視点が欠かせません。当社は本事業の取組みを通じて、社会全体のリテラシー向上に貢献します。

# 目次

1. 開発・実証のサマリ
  1. 開発・実証のサマリ
2. 開発・実証の背景・目的
  1. 開発技術によりアプローチする課題
  2. 開発技術により目指す姿・ゴール
  3. 開発技術により対処可能なユースケース
3. 開発・実証における「対策技術の開発」
  1. 技術開発の全体像
  2. 技術開発の個別詳細
4. 開発・実証における「対策技術の有効性等に関する検証及び調査」
  1. 検証及び調査の全体像
  2. 検証及び調査の個別詳細
5. 開発・実証における「対策技術の社会実装に向けた取組」
  1. 社会実装に向けた取組の全体像
  2. 社会実装に向けた取組の個別詳細
6. 開発・実証における「普及啓発活動への協力」
  1. 普及啓発活動の全体像
  2. 普及啓発活動の個別詳細
7. 開発・実証の課題・展望
  1. 技術開発及び社会実装における課題・展望
8. 開発・実証の実施体制等
  1. 実施体制及び役割分担
  2. 全体スケジュール

## 2-1. 開発技術によりアプローチする課題

### 開発技術によりアプローチする課題（概要）

- 当社は前年度事業において、海外ツールを上回る高精度なディープフェイク対策技術・ツールを開発・実証し、その実用性を検証した。また、研究者やメディア等の有識者との意見交換から、偽・誤情報がSNSユーザーによる生産・拡散・消費プロセスで急速に広がり社会的影響を与える一方で、ユーザーがこうした情報に自律的かつ効果的に対処するための支援の仕組みやツールが十分に普及していない現状が、この偽・誤情報問題の深刻度を高め、対策の喫緊性を増していると強く認識した。
- そこで本事業では、前年度成果を基盤とし、対応範囲をディープフェイクに留まらない多様な偽・誤情報へと拡張した上で、SNSユーザーへの直接的な技術的支援を中核に据えつつ、メディア、官公庁、企業、研究機関などが連携し、訂正情報の発信やリテラシー向上プログラムの提供といった多角的な役割を担う社会全体での対策を推進する。この方針のもと、偽・誤情報の「予防的対応」、「拡散」、拡散後の「訂正」の各段階における課題に対処する技術・ツールを開発・実証することで、健全なデジタル空間の実現に貢献する。

予防的対応（SNSユーザー）

課題1：リテラシーの向上

拡散（SNSユーザー）

課題2：適切な情報検証  
課題3：生成AIへの対応

訂正（メディア、官公庁、企業）

課題4：真偽不明言説の収集・  
訂正情報の周知

## 2-1. 開発技術によりアプローチする課題

### 開発技術によりアプローチする課題（詳細）

#### ・ 課題1：SNSユーザーのICTリテラシー向上における自発的な学習の限界（事前段階）

総務省の調査[1]によれば、9割がICTリテラシーの重要性を認識する一方で、75.3%はリテラシー向上に向けた具体的な取組をしておらず、その理由として、50.9%が「取組み方がわからない」と回答している。この結果は、ウェブ上に様々な学習コンテンツが存在する現状を踏まえると、多くのユーザーが自ら積極的に情報リテラシー向上の学習コンテンツを探し、取組むという自発的な努力に期待することに限界があることを示唆しており、日常生活で自然に学べる機会の創出が、社会全体のICTリテラシー向上には不可欠である。

#### ・ 課題2：SNSユーザーの主体的かつ適切な情報検証を支援するツールの不足（拡散段階）

国際大学GLOCOMの調査[2]によると、多くのSNSユーザーは真偽不明な情報・言説（以下、「真偽不明言説」）を主体的に検証する習慣を持たない（「1次ソース確認」20%、「画像検索」6.7%）。昨今流行するX上のAIファクトチェックは誤判定リスクや根拠先URLが提示されずにユーザーの思考停止を助長するといった懸念も生じている。さらに、主要な生成AIが膨大なプロパガンダ（特定の考えを信じさせるために、都合の良い情報だけを広めること）記事を学習・汚染された事例[3]も報告され、AIの出力を鵜呑みにすることは危険である。このため、SNSユーザーが手軽かつ主体的に情報検証できるツールの普及が急務である。

出典：

[1] 総務省「ICTリテラシー実態調査」、2025年5月

[2] 国際大学GLOCOM「偽・誤情報、ファクトチェック、教育啓発に関する調査研究」、2024年4月

[3] NewsGuard社「A well-funded Moscow-based global 'news' network has infected Western artificial intelligence tools worldwide with Russian propaganda」2025年3月

## 2-1. 開発技術によりアプローチする課題

### 開発技術によりアプローチする課題（詳細）

#### • 課題3：急速に巧妙化する生成AIによるディープフェイクへの対応（拡散段階）

画像、動画、音声の生成AIの品質は急速に高まり続けており、今後、より手軽に誰でもディープフェイクを作成できるようになることで、SNS上にはディープフェイクコンテンツが溢れかえり、社会全体での情報検証コストが増大することは確実である。ディープフェイクの検出技術はその一つの解決策となり得るが、高い精度を維持するためには、低コストかつ高頻度にアップデートする技術の確立が不可欠であり、前年度に開発した高精度なディープフェイク対策技術の強化と継続的な運用技術が求められる。

#### • 課題4：手間のかかる真偽不明言説の収集や広範な訂正情報の周知（訂正段階）

訂正情報対策の第一歩は真偽不明言説の収集だが、膨大なSNS情報から人手で網羅的に収集することは現実的でない。そのため、真偽不明言説を自動かつ大規模に収集し、影響度や緊急度を元に選別（トリアージ）する仕組みが求められる。また、偽・誤情報を一度信じると訂正が困難（誤情報持続効果）との指摘[4]もあり、信じた後のデバンキング（事実に基づき偽・誤情報の誤りを証明し、人々の誤解を解くこと）だけでなく信じる前の情報に接するタイミングで介入するアプローチも重要である

出典：

[4] 田中優子・犬塚美輪・藤本和則、「誤情報持続効果をもたらす心理プロセスの理解と今後の展望：誤情報の制御に向けて」、29 巻 3 号 p. 509-527 (2022)

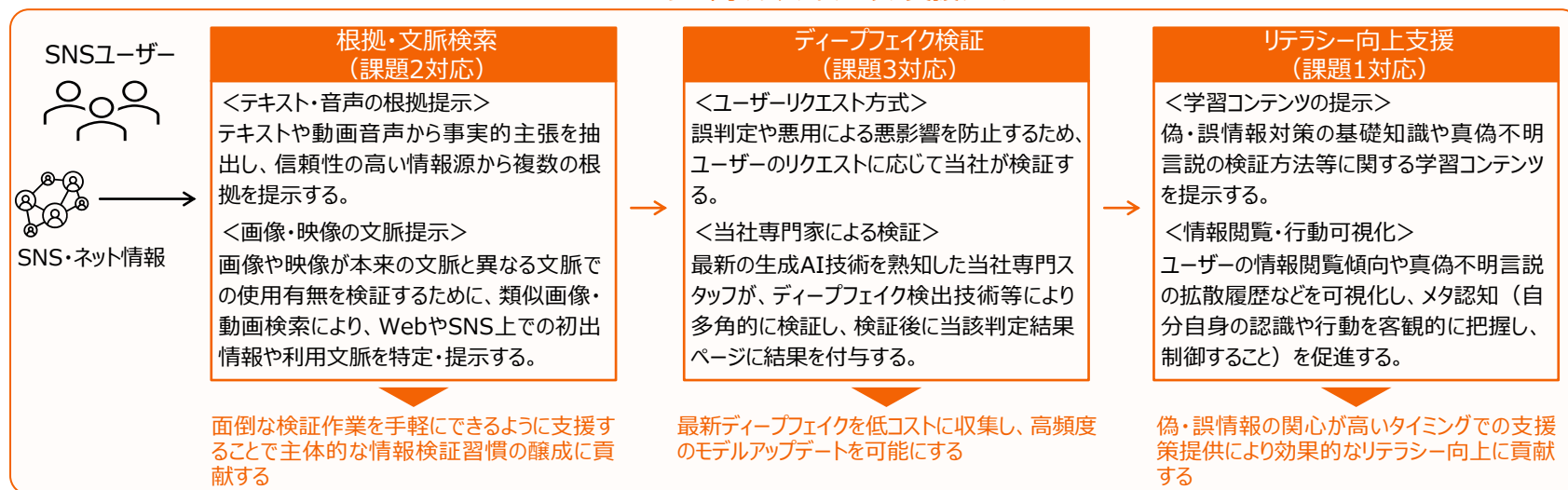
## 2-2. 開発技術により目指す姿・ゴール

### 開発技術を通して目指す姿・ゴール

- 本事業では、前述した4つの課題解決に貢献する『SNSユーザー向けファクトチェック支援ツール』及び『訂正主体向けSNSモニタリングツール』を開発し、その有効性やユースケースを検証・調査する

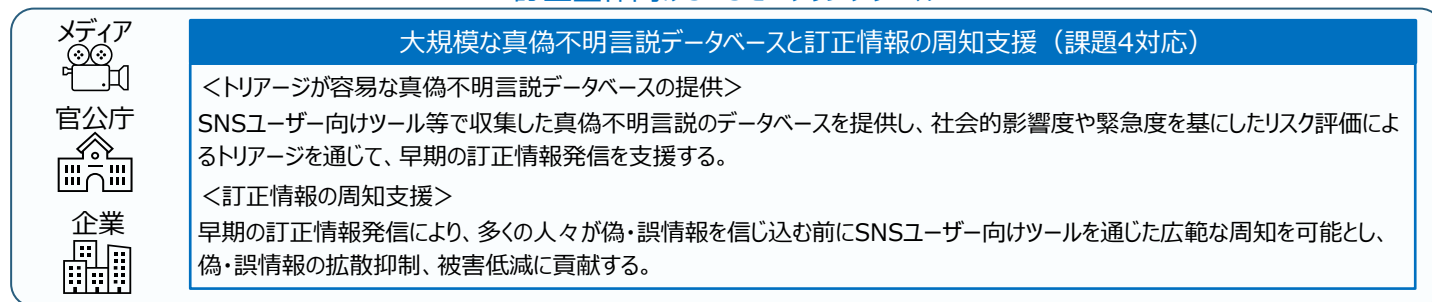
#### SNSユーザーを中心とした対策の全体像

#### SNSユーザー向けファクトチェック支援ツール



#### 訂正主体向けSNSモニタリングツール

真偽が不明の情報が確認された

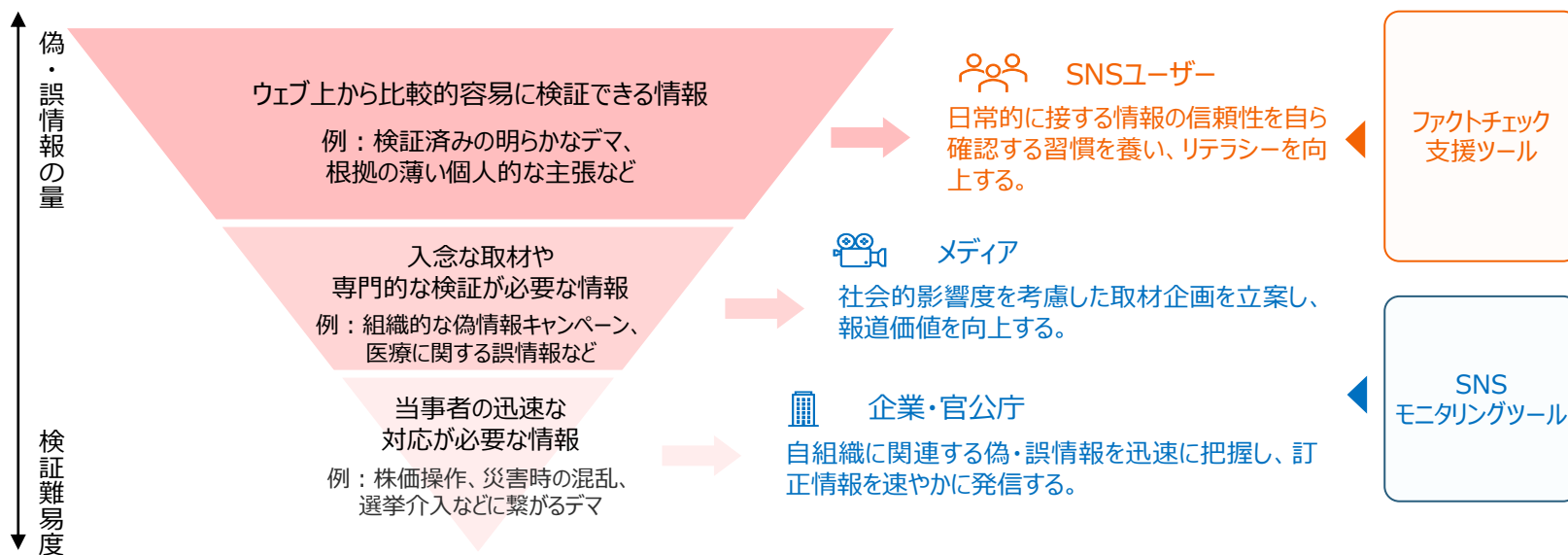


## 2-2. 開発技術により目指す姿・ゴール

### 開発技術を通して目指す姿・ゴール

- 偽・誤情報対策における構造的な課題である『情報を発信する側』が『検証する側』（メディア等）を数で圧倒的に上回るという非対称性に対応するためには、検証難易度に応じた社会的な役割分担が重要である。そこで、本事業では、①検証が比較的容易な情報はSNSユーザーがメディア・リテラシーを高め、自ら対応するとともに安易な拡散を抑制することを中核の戦略としつつも②専門的なスキルや多くのリソースを必要とする情報はメディアが担い、③速やかな対応が求められる株価操作や災害関連のデマ、選挙介入などはそれらの当事者である企業や官公庁による訂正情報の発信を行う。このような役割分担を促進するツールを提供することで、社会全体での偽・誤情報の抑制に貢献する。

提案ツールのユースケース分類（左図）と検証主体別の期待される効果（右図）



## 2-3. 開発技術により対処可能なユースケース

### 開発技術により対処可能なユースケース（概要）

- モダリティ毎に対処可能なユースケースと適用する開発技術を整理した

提案ツールによる偽・誤情報の検証方法		
文章	真偽不明言説	要素技術2.根拠情報提示技術：主張に対する根拠情報を確認・検証
	AIファクトチェック	要素技術2.根拠情報提示技術：判定結果に対する根拠情報を確認・検証
画像	文脈操作	要素技術3.出所・文脈提示技術：本来とは異なる文脈で画像を利用しているかを確認・検証
	チープフェイク	要素技術3.出所・文脈提示技術：変更前の画像を特定することで検証
	ディープフェイク	要素技術1.ディープフェイク検出技術：判定結果を参考に検証 要素技術3.出所・文脈提示技術：AI変更の場合、変更前の画像を特定・検証
動画	文脈操作	要素技術3.出所・文脈提示技術：本来とは異なる文脈で動画を利用しているかを確認することで検証
	真偽不明言説	要素技術2.根拠情報提示技術：動画音声の文字起こし結果における主張に対する根拠情報を確認・検証
	ディープフェイク	要素技術1.ディープフェイク検出技術：判定結果を参考に検証 要素技術3.出所・文脈提示技術：AI変更の場合、変更前の動画を特定・検証
音声	真偽不明言説	要素技術2.根拠情報提示技術：音声の文字起こし結果における主張に対する根拠情報を確認・検証
	ディープフェイク	要素技術1.ディープフェイク検出技術：判定結果を参考に検証

※各要素技術については 3.開発・実証における「対策技術の開発」にて詳細に説明する

※SNSユーザー向けツールと訂正主体向けツールの両方に上記技術は実装済みであるが、SNSユーザー向けツールについてはユーザーの受容性や認知的負荷を考慮し、まずは要素技術2のみを搭載する形式での一般公開としている。

## 2-3. 開発技術により対処可能なユースケース

### 開発技術により対処可能なユースケース（文章）

- 文章に対しては、「根拠情報提示技術」を活用し、SNS上の真偽不明な言説やAIによるファクトチェック結果に対し、ユーザー自身がその信憑性を容易に検証・判断するための根拠情報を収集・提示する。

ユースケース	想定ニーズ	本事業での対応方法
真偽不明言説	<ul style="list-style-type: none"> <li>SNSを閲覧・視聴する中で、「これは本当？」という情報に接したときに手軽にチェックしたい</li> </ul>	<ul style="list-style-type: none"> <li>根拠情報提示技術を用いて、ユーザーが当該情報の真偽を容易に検証するための情報を収集・提示する</li> </ul>
AIファクトチェック	<ul style="list-style-type: none"> <li>SNSで流行しているAIファクトチェックの結果をみても、その根拠情報のURL示されていないため、そのチェックまで自動でしたい</li> </ul>	<ul style="list-style-type: none"> <li>根拠情報提示技術を用いて、ユーザーがAIファクトチェック結果を容易に検証するための根拠情報を収集・提示する</li> </ul>

## 2-3. 開発技術により対処可能なユースケース

### 開発技術により対処可能なユースケース（画像）

- 画像に対しては、「出所・文脈提示技術」と「ディープフェイク判定技術」を活用し、画像の文脈操作や加工、生成AIによる偽造の疑いに対し、初出情報や作成痕跡を提示することで、ユーザーによる真偽検証を支援する。

ユースケース	想定ニーズ	本事業での対応方法
文脈操作	<ul style="list-style-type: none"> <li>本来の文脈とは異なる文脈で拡散された画像であるかをチェックしたい（EX: 災害時に過去の災害の画像が拡散される）</li> </ul>	<ul style="list-style-type: none"> <li>出所・文脈提示技術を活用して、初出情報を検索・提示する</li> </ul>
チープフェイク	<ul style="list-style-type: none"> <li>切り抜きや加工により改竄された画像であるかをチェックしたい</li> </ul>	<ul style="list-style-type: none"> <li>出所・文脈提示技術を活用して、初出情報を検索・提示する</li> </ul>
ディープフェイク	<ul style="list-style-type: none"> <li>生成AIによって作られたコンテンツであるかをチェックしたい</li> </ul>	<ul style="list-style-type: none"> <li>ディープフェイク判定技術を活用して、ユーザーに生成AIによる作成・加工の痕跡があるかを提示する</li> </ul>

## 2-3. 開発技術により対処可能なユースケース

### 開発技術により対処可能なユースケース（動画）

- 動画に対しては、「出所・文脈提示」「根拠情報提示」「ディープフェイク判定」の各技術を活用し、動画の文脈操作や不確かな言説、AI生成の疑いに対し、検証情報をユーザーへ提示する。

ユースケース	想定ニーズ	本事業での対応方法
文脈操作	<ul style="list-style-type: none"> <li>• 本来の文脈とは異なる文脈で拡散された動画であるかをチェックしたい（EX: 災害時に過去の災害の動画が拡散される）</li> </ul>	<ul style="list-style-type: none"> <li>• 出所・文脈提示技術を活用して、初出情報を検索・提示する</li> </ul>
真偽不明言説	<ul style="list-style-type: none"> <li>• SNSを閲覧・視聴する中で、「これは本当？」という情報に接したときに手軽にチェックしたい</li> </ul>	<ul style="list-style-type: none"> <li>• 根拠情報提示技術を用いて、ユーザーが当該情報の真偽を容易に検証するための情報を収集・提示する</li> </ul>
ディープフェイク	<ul style="list-style-type: none"> <li>• 生成AIによって作られたコンテンツであるかをチェックしたい</li> </ul>	<ul style="list-style-type: none"> <li>• ディープフェイク判定技術を活用して、ユーザーに生成AIによる作成・加工の痕跡があるかを提示する</li> </ul>

## 2-3. 開発技術により対処可能なユースケース

### 開発技術により対処可能なユースケース（音声）

- 音声に対しては、「根拠情報提示技術」と「ディープフェイク判定技術」を併用し、音声による真偽不明な言説や、生成AIによる声のなりすまし等に対し、その裏付け情報や生成有無を提示してユーザーの検証を支援する。

ユースケース	想定ニーズ	本事業での対応方法
真偽不明言説	<ul style="list-style-type: none"><li>• SNSを閲覧・視聴する中で、「これは本当？」という情報に接したときに手軽にチェックしたい</li></ul>	<ul style="list-style-type: none"><li>• 根拠情報提示技術を用いて、ユーザーが当該情報の真偽を容易に検証するための情報を収集・提示する</li></ul>
ディープフェイク	<ul style="list-style-type: none"><li>• 生成AIによって作られたコンテンツであるかをチェックしたい</li></ul>	<ul style="list-style-type: none"><li>• ディープフェイク判定技術を活用して、ユーザーに生成AIによる作成・加工の痕跡があるかを提示する</li></ul>

# 目次

1. 開発・実証のサマリ
  1. 開発・実証のサマリ
2. 開発・実証の背景・目的
  1. 開発技術によりアプローチする課題
  2. 開発技術により目指す姿・ゴール
  3. 開発技術により対処可能なユースケース
3. 開発・実証における「対策技術の開発」
  1. 技術開発の全体像
  2. 技術開発の個別詳細
4. 開発・実証における「対策技術の有効性等に関する検証及び調査」
  1. 検証及び調査の全体像
  2. 検証及び調査の個別詳細
5. 開発・実証における「対策技術の社会実装に向けた取組」
  1. 社会実装に向けた取組の全体像
  2. 社会実装に向けた取組の個別詳細
6. 開発・実証における「普及啓発活動への協力」
  1. 普及啓発活動の全体像
  2. 普及啓発活動の個別詳細
7. 開発・実証の課題・展望
  1. 技術開発及び社会実装における課題・展望
8. 開発・実証の実施体制等
  1. 実施体制及び役割分担
  2. 全体スケジュール

## 3-1. 技術開発の全体像

### 技術開発に係る取組・成果の全体像

本技術開発ではSNSでの偽・誤情報に対抗するための要素技術の開発として以下の5つに取り組んだ。

#### <要素技術 1. 画像・映像・音声のディープフェイク検出の運用・高度化技術>

- 前年度開発した高精度な画像・動画・音声ディープフェイク検出エンジンを改善・強化するとともに、進化し続ける生成 AI に対応するために低コストで継続的な改善ができる運用システムを構築し、DF検出精度93.1%を達成

#### <要素技術 2. Deep Research 技術による信頼性評価付き根拠情報検索技術>

- SNS 上のテキスト、動画、音声の真偽不明言説の検証に不可欠である信頼性の高い複数の根拠情報に容易にアクセスできる技術を開発し、実際に流通した偽・誤情報に対する根拠情報提示において、正しい根拠を提示する割合（適合率）92.9%を達成

#### <要素技術 3. 画像・動画に対する出所及び文脈の検証技術>

- SNS 上に投稿された画像動画に対して、リバースサーチである類似画像検索や類似動画検索を実施し、Web 上や他の SNS プラットフォーム上での初出情報や利用文脈（どのような記事や投稿と共に使われてきたか）を特定・提示する技術を開発し、SNS 上で実際に流通した改変・流用画像/動画に対し、正しい初出情報・利用文脈を提示する割合（適合率）90%を達成

#### <要素技術 4. 情報接種・行動の分析技術>

- SNS ユーザーのタイムラインで流れてくる情報の偏りの分析や、「いいね」や「リポスト」などのエンゲージメント行動をした情報に偽・誤情報が含まれていたかなど、自らの情報環境をわかりやすく振り返る機会を提供する技術を開発し、ラベリングデータで情報接種分析の結果に関して QWK(Quadratic Weighted Cohen's  $\kappa$ ) 0.62 を達成する。※QWK：多クラス分類の評価指標の1つで、混同行列に重みを付けて評価する指標

#### <要素技術 5. 真偽不明言説のリスクアセスメント・拡散予測技術>

- 真偽不明言説に対して、リスク評価技術によるトリアージ支援と拡散予測技術による早期の対抗策実施を支援する技術を開発し、ツールが算出したリスクスコアに関して QWK 0.84 を達成

## 3-2. 技術開発の個別詳細

### 要素技術 1. 画像・映像・音声のディープフェイク検出の運用・高度化技術 – ディープフェイク検出エンジン

#### • 目的

- 生成 AI の急速な発展に対応するために、最新の生成AIで作られたDFデータセットを構築し、それをを用いて昨年度開発したDF検出エンジンをファインチューニング

#### • 取組内容

- 実際にSNSで流通し、対応が求められるリアルなディープフェイクデータの収集
- 昨年度構築したディープフェイク生成基盤を拡張子、NanoBanana Proを始めとする最新の生成でディープフェイクデータセットを構築
- 昨年度開発したSYNTHETIQ VISION を基盤とした検出エンジンの検知精度、頑健性及び、ファインチューニングによる精度向上

#### • 得られた成果

- 最新の生成AIを含む包括的なディープフェイクデータセットを構築
- 昨年度開発したディープフェイク検出エンジンを最新のDFデータセットの一部を用いてファインチューニングし、最新のDFの評価データにおいてもディープフェイクの検出精度93.1%を実現。

#### • KPI

- 最新の生成 AI で作成された多様なディープフェイクを含む評価データセットに対する検出精度 93.1%を達成 (KPI : 90%)

## 3-2. 技術開発の個別詳細

### 要素技術 1. 画像・映像・音声のディープフェイク検出の運用・高度化技術 – ディープフェイク検出エンジン表と図

今回構築したディープフェイクデータセットの内訳

生成AIサービス名	枚数
Midjourney	10434
Higgsfield	6534
Akool	4787
Nano Banana Pro	3921
Canva AI	2159
Seadream	2079
合計	29914

作成したフェイク画像データの例



## 3-2. 技術開発の個別詳細

### 要素技術 1. 画像・映像・音声のディープフェイク検出の運用・高度化技術 – MLOps

- **目的**
  - 生成 AI の急速な発展に対応するために、機械学習技術の開発から運用、改善までの一連のフローを効率化する MLOps (Machine Learning Operations) パイプラインを構築
- **取組内容**
  - データ収集、再学習、評価までを行う MLOps パイプラインの全体アーキテクチャ設計
  - 検出エンジンの性能をテストし、弱点を評価するテストモジュールの設計・開発
  - チューニングデータ生成モジュールの設計・開発
  - パイプラインの構築と運用テスト
- **得られた成果**
  - 最新の生成 AI 技術によって作られたディープフェイクに対しても短期間の再学習で自動的に検知できる MLOps パイプラインを構築することができた

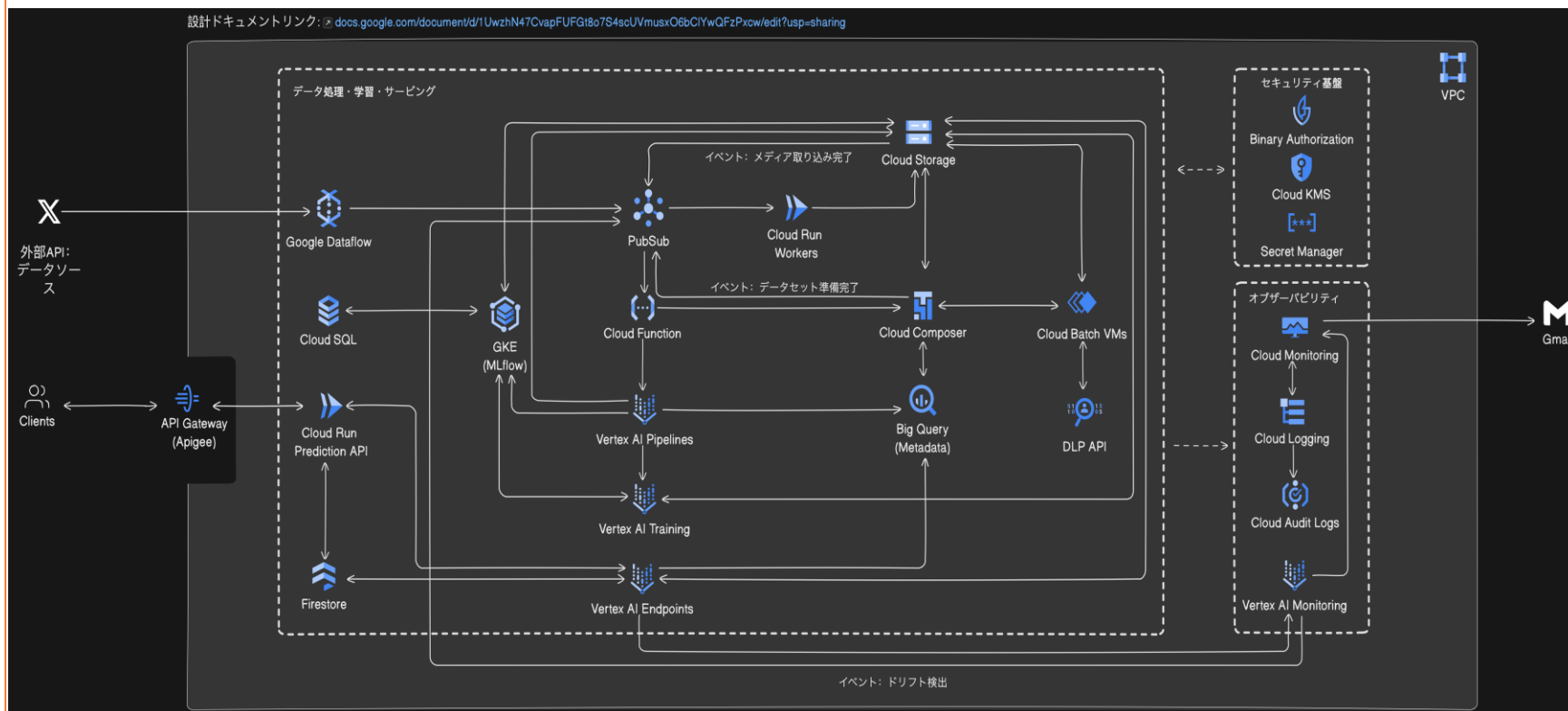
## 3-2. 技術開発の個別詳細

### 要素技術 1. 画像・映像・音声のディープフェイク検出の運用・高度化技術 – MLOps

• **特徴**

- Vertex AI Monitoringにてモデル・データ品質を監視しており、流通しているディープフェイクデータの分布の変化やディープフェイク検知AIモデルの精度劣化を常時監視し、リアルタイムでのモデルアップデートのための基盤を構築

#### 構築したMLOpsアーキテクチャ



## 3-2. 技術開発の個別詳細

### 要素技術 2. Deep Research 技術による信頼性評価付き根拠情報検索技術

#### • 目的

- SNS 上のテキスト、動画、音声の真偽不明言説の検証に不可欠である信頼性の高い複数の根拠情報に容易にアクセスできる技術を開発

#### • 取組内容

- 事実的主張抽出モジュールの設計・開発
- 情報源信頼性評価モジュールの設計・開発
- Deep Research型（情報を掘り下げ・繰り返し検索しながら理解を深める方式）の根拠情報検索モジュールの設計・開発
- 複数の根拠を統合して判断するの設計・実装・評価

#### • 得られた成果

- SNS ユーザーが真偽不明言説を検証する際に、信頼できる支持・反証両面の根拠情報への容易なアクセスを実現し、検証作業を大幅に効率化した。さらに、ユーザー自身が根拠情報に基づき主体的に真偽を判断の支援になることが期待される。

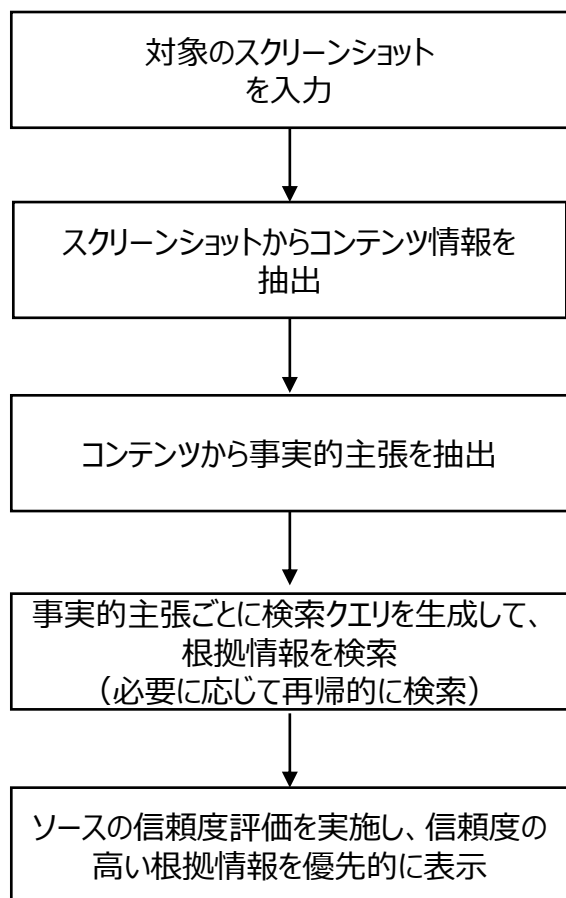
#### • KPI

- SNS 上で実際に流通した偽・誤情報に対する根拠情報提示において、正しい根拠を提示する割合（適合率）**92.9%**を達成

## 3-2. 技術開発の個別詳細

### 要素技術 2. Deep Research 技術による信頼性評価付き根拠情報検索技術 – UI / ワークフロー

#### ワークフロー



#### 信頼性評価付き根拠情報検索のUI



## 3-2. 技術開発の個別詳細

### 要素技術 3. 画像・動画に対する出所及び文脈の検証技術

- **目的**
  - SNS 上の画像や動画コンテンツの初出情報及び過去にどのような文脈で利用されてきたかを特定・追跡し、検証に必要な情報を提示する技術を開発する。
- **取組内容**
  - **リバーシメージサーチの拡張**
    - テロップなど加工箇所を除いた領域抽出技術の設計・開発
    - シーンチェンジや内容変化度を考慮したキーフレーム抽出アルゴリズムの設計・開発
  - **VLM による視覚情報を利用した初出情報・過去利用文脈の特定技術**
    - 動画内のオブジェクト認識や行動認識を行い、動画内容の説明テキスト生成技術の設計・開発
    - 意味的テキスト検索機能の設計・開発
  - **ユーザビリティの高いアウトプットの表示形式の設計**
    - 画像・動画が含まれる「信頼度の高い」情報源を検索し、ユーザビリティが高いUIで表示
- **得られた成果**
  - 画像や動画に対するリバーシサーチ（類似画像・動画検索）と VLM（Vision Languageモデル; 視覚情報と言語情報の両方を使ったモデル）による視覚的理解を基にした検索技術により、本来の文脈で画像や動画が使われているかを迅速に検証することが可能になった
- **KPI**
  - SNS 上で実際に流通した改変・流用画像/動画に対し、正しい初出情報・利用文脈を提示する割合（適合率）**90%**を達成



## 3-2. 技術開発の個別詳細

### 要素技術 4. 情報接種・行動の分析技術

#### • 目的

- SNS ユーザーのタイムラインで流れてくる情報の偏りの分析や、「いいね」や「リポスト」などのエンゲージメント行動をした情報に偽・誤情報が含まれていたかなど、自らの情報環境をわかりやすく振り返る機会を提供する技術を開発する。

#### • 取組内容

- 個人情報に配慮したデータ連携技術の設計・開発
- 接種情報の分析技術の設計・開発
- エンゲージメント行動の分析技術の設計・開発
- 可視化インターフェースの設計・開発

#### • 得られた成果

- SNS ユーザーに対して、情報接種の偏りやエンゲージメント行動における真偽不明言説の拡散履歴などの可視化により、情動的健康的意識するきっかけと行動変容に繋がる機会を提供することができた

#### • KPI

- ラベリングデータで情報接種分析の結果に関して **QWK(Quadratic Weighted Cohen's  $\kappa$ )0.615** を達成。※QWK：多クラス分類の評価指標の 1 つで、混同行列に重みを付けて評価する指標

## 3-2. 技術開発の個別詳細

### 要素技術 4. 情報接種・行動の分析技術 UI/ワークフロー

Xアカウントを連携して、自分が反応したポストを確認



ユーザーに対して、情報接種の偏りやエンゲージメント行動における真偽不明言説の拡散履歴などの可視化により、情報的健康を意識するきっかけと行動変容に繋がる機会を提供することができた

## 3-2. 技術開発の個別詳細

### 要素技術 5. 真偽不明言説のリスクアセスメント・拡散予測技術

#### ・ 目的

- ・ 真偽不明言説に対して、リスク評価技術によるトリアージ支援と拡散予測技術による早期の対抗策実施を支援する技術を開発する。

#### ・ 取組内容

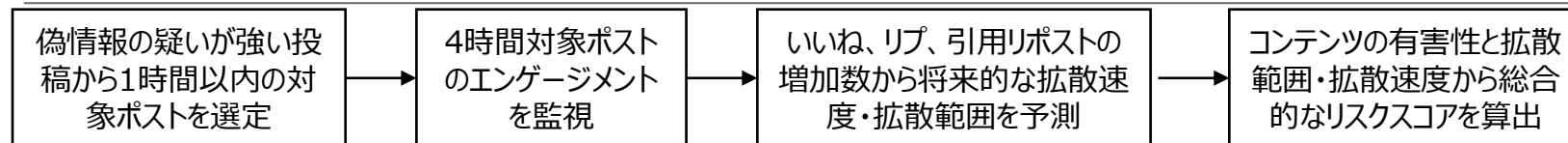
##### ・ リスクスコア算出モジュール群の開発

- ・ 拡散速度、影響範囲等の各リスク指標算出アルゴリズムの開発・実装
- ・ 各指標を統合し、総合的なリスクスコアリングモデルの設計・実装

##### ・ 拡散ネットワークの高度分析と将来予測機能の実装

- ・ 収集したネットワークデータの基づき、拡散構造を分析し、可視化機能を実装
- ・ 過去の蓄積データから拡散拡大のきっかけとなるブローカー特定・将来のアウトブレイク予測アルゴリズムの開発・実装

リスクスコア算出モジュールのワークフロー



#### ・ 得られた成果

- ・ 訂正主体において、真偽不明言説に対し、対応すべき優先度を即時かつ的確に判断するための高度なリスク評価技術と実用的なトリアージシステムを確立し、高リスク事案の特定率向上や検知時間改善を実現

#### ・ KPI

- ・ ツールが算出したリスクスコアに関して **QWK 0.867** を達成

# 目次

1. 開発・実証のサマリ
  1. 開発・実証のサマリ
2. 開発・実証の背景・目的
  1. 開発技術によりアプローチする課題
  2. 開発技術により目指す姿・ゴール
  3. 開発技術により対処可能なユースケース
3. 開発・実証における「対策技術の開発」
  1. 技術開発の全体像
  2. 技術開発の個別詳細
4. 開発・実証における「対策技術の有効性等に関する検証及び調査」
  1. 検証及び調査の全体像
  2. 検証及び調査の個別詳細
5. 開発・実証における「対策技術の社会実装に向けた取組」
  1. 社会実装に向けた取組の全体像
  2. 社会実装に向けた取組の個別詳細
6. 開発・実証における「普及啓発活動への協力」
  1. 普及啓発活動の全体像
  2. 普及啓発活動の個別詳細
7. 開発・実証の課題・展望
  1. 技術開発及び社会実装における課題・展望
8. 開発・実証の実施体制等
  1. 実施体制及び役割分担
  2. 全体スケジュール

## 4-1. 検証及び調査の全体像

### 検証及び調査に係る取組・成果の全体像

- SNSユーザー向けツールおよび訂正主体向けツールを開発し、その有効性、活用方法、普及可能性を検証するため、以下の3段階の検証を実施する。各検証ではその目的に対応するビジネス面でのKPIを設定することで、達成度を定量的に評価するとともに、アンケート、インタビューを通じて、技術面、運用面、倫理面等の社会実装に向けた課題を抽出し、その解決策を調査・検討する

検証フェーズ	検証目的
① SNSユーザー向けツール クローズドアルファ版テスト	<ul style="list-style-type: none"> <li>● 開発したファクトチェック支援機能が設計通りに機能するかを検証するとともに、ツール利用によるユーザーの情報リテラシー向上および意識変容の効果を多角的に評価する。</li> </ul>
② SNSユーザー向けツール パブリックベータ版テスト	<ul style="list-style-type: none"> <li>● SNS上の実環境においてツールがどのように受容・活用されるかの実態を定量的に把握し、幅広い一般ユーザーへの普及可能性および社会実装に向けたポテンシャルを評価する。</li> </ul>
③ 訂正主体向けツール テスト検証	<ul style="list-style-type: none"> <li>● 報道機関のポテンシャルユーザーと共同で実務への適合性を実証し、プロフェッショナルのワークフローにおけるツール導入の有用性と運用の実現性を検証する。</li> </ul>

## 4-2. 検証及び調査の個別詳細

### 検証 1: SNS ユーザー向けツールのクローズドアルファ版テスト（実施概要）

カテゴリ	内容
目的	主要な機能に限定して試作したアルファ版のSNSユーザー向けツールの①ファクトチェック支援の有用性、②情報リテラシー向上効果を評価し、一般ユーザーでの利用を見据えた改善点を抽出する。
対象	50～100名程度のモニター参加者（非公開で実施） モニター参加者は年齢、性別、SNS利用頻度等の多様性を考慮して選定する。なお、参加者には事前に実験の目的、内容、データ利用について十分な説明・同意を取得した上で実施する。
期間	2025年12月～2026年2月（3ヶ月間）
実施内容	<ul style="list-style-type: none"> <li>モニター参加者が実際にSNS（X・YouTube）を利用する中で積極的に活用いただく</li> <li>利用ログ（リクエスト頻度、閲覧情報種別、根拠情報クリック率等）を匿名化した上で収集する</li> <li>改善サイクルを回すために、2週間程度を1タームとして、上記を複数回実施する方式とする</li> <li>実験前に、現状の情報リテラシーを測るためのアンケートを実施</li> <li>ツールを利用し感じた評価や改善点をアンケート形式で収集</li> <li>実験後に、情報リテラシーの変化を測るためのアンケートを実施</li> <li>アンケート例（各設問ではそれぞれ7段階で評価） <ul style="list-style-type: none"> <li>SNSの投稿を見たとき、根拠となる情報を確認してから判断することが多い</li> <li>複数の根拠情報を見ることで、情報の信頼性を考えることができると思う</li> <li>いいね・リポストをする前には、慎重な判断が必要だと思う</li> </ul> </li> </ul>
KPI	<ul style="list-style-type: none"> <li>コアユーザーの満足度評価：10段階中7以上（①ファクトチェック支援の有用性）</li> <li>リテラシー評価：ツール利用前後で有意な情報リテラシー向上（②情報リテラシー向上効果）</li> </ul> <p>※ サンプルサイズやアンケートの学習効果等を専門家とともに考慮して設計</p>

## 4-2. 検証及び調査の個別詳細

### 検証 1: SNS ユーザー向けツールのクローズドアルファ版テスト（結果）

- コアユーザーの満足度評価：10段階中7以上（①ファクトチェック支援の有用性）
  - 達成：総合評価7.81
  - 分析
    - ポジティブ意見（抜粋）
      - 物事の真偽判断能力を向上させると思う
      - 自分とは別の客観的な判断基準として使えるから
    - ネガティブ意見（抜粋）
      - アプリを信頼することで逆に自己判断能力は下がるのではないか
      - 根拠情報の精度がもっと上がれば良いと思う
- リテラシー評価：ツール利用前後で有意な情報リテラシー向上（②情報リテラシー向上効果）
  - 達成
  - 分析
    - 59名を対象とした対応のあるt検定の結果、アプリ使用前後で統計的に有意な差が認められた。平均スコアは作業前66.4点、作業後69.7点へと+3.3点向上し、効果量は中程度である。

項目	検証結果
p値	0.0002
サンプル数	59人
変化量	+3.3点
効果量	0.52（中程度）

## 4-2. 検証及び調査の個別詳細

### 検証 2: SNS ユーザー向けツールのパブリックベータ版テスト（実施概要）

カテゴリ	内容
目的	検証1を踏まえて完成度を高めたベータ版のSNSユーザー向けツールを一般公開し、①普及ポテンシャルの評価、②ツールの評価について検証し、社会実装に向けた技術面・運用面・倫理面の課題を明確化し、対応方針を策定する。
対象	一般SNSユーザー（一般公開もしくはウェイトングリスト方式による提供を想定）
期間	2026年1月～2026年2月（2ヶ月間）
実施内容	<ul style="list-style-type: none"><li>● 一般SNSユーザーが実際にSNS（X・YouTube）を利用する中で活用いただく</li><li>● 利用ログ（リクエスト頻度、閲覧情報種別、根拠情報クリック率等）を匿名化した上で収集する</li><li>● 実際の活用方法（真偽不明言説のファクトチェック、AIファクトチェック出力の根拠チェック、コミュニティノート作成など）を分析する</li><li>● 誤用のモニタリングやユーザーの反応を分析する</li><li>● 専用の問い合わせフォームを設置し、ユーザーからの質問、バグ報告、意見・要望に対応する</li></ul>
KPI	<ul style="list-style-type: none"><li>● マンスリーアクティブユーザー：500アカウント以上（①普及ポテンシャルの評価）</li><li>● SNS上での口コミ評価：ポジティブ評価が70%以上（②ツールの評価）</li></ul>

## 4-2. 検証及び調査の個別詳細

### 検証 2: SNS ユーザー向けツールのパブリックベータ版テスト (結果)

- マンスリーアクティブユーザー：500アカウント以上 (①普及ポテンシャルの評価)
  - 未達成：2/24(火)時点で209
  - 分析
    - アプリのリリース時点でのプレスリリースと毎日新聞社による記事掲載によって一定数のユーザー獲得には至ったが、選挙期間との重複などにより、ユーザーの認知度向上にはまずまずの結果となった。
    - 検証1の結果より、情報リテラシーの向上には一定の有意差が認められるため、各メディアと連携し、各種PRを促進し、ユーザー数獲得、認知度向上に取り組むことを次の重点的施策として取り組む。
- SNS上での口コミ評価：ポジティブ評価が70%以上 (②ツールの評価)
  - 達成：76.3% (テスターから収集した結果より分析)
  - ユーザーからの意見 (抜粋)
    - ポジティブ
      - 自分の検索力では辿りつかないようなソースを提示してくれる
      - 自分で調べる機会を作り、それを習慣化できるため
      - スクリーンショットだけで情報の真偽についてチェックできて、気軽に使えて便利
    - ネガティブ
      - 分析時間が長い
      - 真偽を確かめようと思う習慣が無い
  - 分析
    - 全体的にポジティブ傾向が見られるものの、アプリ自体への信頼性や情報を検証する習慣が無いユーザーにとっては魅力を感じにくい点が課題である。上述の通り、各メディアと連携しアプリの認知度向上を図りつつ、アプリのUI/UX向上にも取り組み、コアユーザーだけでなく、より幅広い層に受けられるアプリとなることを次の目標として取り組む。

## 4-2. 検証及び調査の個別詳細

### 検証 3: 訂正主体向けツールのテスト検証（実施概要）

カテゴリ	内容
目的	メディア、官公庁、企業などの訂正主体におけるヒアリングやテスト利用を通じて、①ツールのユーザービリティ、②実業務利用に資するユースケースの調査・検討を行い、社会実装に向けたツールのあるべき姿やターゲットユーザーを明確にする。
対象	訂正主体（メディア、官公庁、企業）3～5社程度
期間	2025年12月～2026年2月（3ヶ月）
実施内容	<ul style="list-style-type: none"><li>● 検証前に仕様・活用法などに関して意見交換会を実施する</li><li>● 偽・誤情報のモニタリングや検証業務の中で、ツールをテスト利用いただく</li><li>● 定期的なフィードバックの収集を通じて、ツール利用による真偽不明言説の収集や訂正情報の発信・周知の業務利用におけるユーザービリティや利用プロセスを検証する</li></ul>
KPI	<ul style="list-style-type: none"><li>● ユーザービリティ評価：10段階中7以上（①ユーザービリティ）</li><li>● ユースケースの明確化：3つ以上（②ユースケース）</li></ul>

## 4-2. 検証及び調査の個別詳細

### 検証 3: 訂正主体向けツールのテスト検証（実施概要）

#### 訂正主体向けツールの表示例



分析結果の表示例。  
根拠情報となる記事を提示し、  
それぞれの主張に対して説明  
している。

## 4-2. 検証及び調査の個別詳細

### 検証 3:訂正主体向けツールのテスト検証（結果）

- 実施内容
  - 報道機関に約2週間の期間中、訂正主体向けSNSモニタリングツールを自由に使用いただき、各2回のヒアリングを実施した。
- 分析
  - ユーザービリティ評価：10段階中7以上（①ユーザービリティ）
    - 達成：10段階中7
    - 分析
      - ヒアリングの結果、両社より実際の業務フローの中で活用できる姿が想像できる、との評価を獲得
  - ユースケースの明確化：3つ以上（②ユースケース）
    - 3つ以上の明確化が完了。本資料には抜粋して3件記載

#### カテゴリ

#### ユースケース

検証対象の早期探知と優先順位付け

急騰ワードの自動検知とアラート

- 特定の言説やハッシュタグの急激な拡散を検知し、記者が人力で行っている検索・監視作業を自動化する。

拡散構造の多角的分析

ネットワークグラフによる属性把握

- 拡散に関与しているアカウント群の繋がりや属性を可視化し、情報の出所や拡散のキーマンを特定する。

高度なアドホック分析

AIを用いた自然言語によるデータ抽出

- 専門的な知識がなくても、AIへの指示によって、真偽不明言説データベースから特定の観点に基づくデータ抽出や可視化を実行する。

# 目次

1. 開発・実証のサマリ
  1. 開発・実証のサマリ
2. 開発・実証の背景・目的
  1. 開発技術によりアプローチする課題
  2. 開発技術により目指す姿・ゴール
  3. 開発技術により対処可能なユースケース
3. 開発・実証における「対策技術の開発」
  1. 技術開発の全体像
  2. 技術開発の個別詳細
4. 開発・実証における「対策技術の有効性等に関する検証及び調査」
  1. 検証及び調査の全体像
  2. 検証及び調査の個別詳細
5. 開発・実証における「対策技術の社会実装に向けた取組」
  1. 社会実装に向けた取組の全体像
  2. 社会実装に向けた取組の個別詳細
6. 開発・実証における「普及啓発活動への協力」
  1. 普及啓発活動の全体像
  2. 普及啓発活動の個別詳細
7. 開発・実証の課題・展望
  1. 技術開発及び社会実装における課題・展望
8. 開発・実証の実施体制等
  1. 実施体制及び役割分担
  2. 全体スケジュール

## 5-1. 社会実装に向けた取組の全体像

### 社会実装に係る取組・成果の全体像

- これらのターゲット市場における円滑な社会実装を推進するため、本事業期間中に以下の3つの取り組みを実施した。

#### 1. ニーズの深掘りと事業計画の精緻化

- 目的：ビジネスモデル確立

##### ヒアリング実施

ユーザー候補企業に対策の課題や予算感をヒアリングし、ニーズを深掘りする

##### 事業モデルの検討

ターゲット別の最適なサービス提供モデルを設計する

#### 2. 戦略的パートナーシップの構築

- 目的：エコシステム形成による社会実装の加速

ソーシャルリスニング事業者

偽・誤情報対策  
コンサル事業者

PR・IR支援  
事業者

協業  
モデル検討

#### 3. グローバル展開の検討

- 目的：国際プレゼンス、競争力向上

##### ツールの多言語対応

英語などの主要言語へ拡大し、海外展開を見据えてシステム設計で開発する

##### 海外動向の調査

米国やイスラエルのスタートアップ企業の動向や業界標準、法規制等を調査・分析する

## 5-2. 社会実装に向けた取組の個別詳細

### 1. ニーズの深掘りと事業計画の精緻化

- 自治体および報道機関に対して、ヒアリングを実施し、ニーズを深掘りした

課題カテゴリ	自治体	報道機関
体制・リソースに関する課題	<ul style="list-style-type: none"> <li>SNS担当が実質1名等、情報対策まで手が回らない</li> <li>専門部署・マニュアルが存在せず、所管課ごとのアドホック対応</li> <li>事案発生時の対処法・判断基準が定まっていない</li> </ul>	<ul style="list-style-type: none"> <li>検証担当が記者1名に属人化し、業務負荷が限界に達している</li> <li>偽情報対策に特化した組織体制がなく、各記者が個別判断で対応</li> <li>ファクトチェック対象の選定理由に客観的根拠を示しにくい</li> </ul>
情報の検知・判断に関する課題	<ul style="list-style-type: none"> <li>拡散後に訂正情報を出しても届かず、收拾がつかない</li> <li>何をもって「偽情報」とするか基準・責任所在が不明確</li> <li>AI生成画像の判別がつかず、現場確認に手間取る</li> </ul>	<ul style="list-style-type: none"> <li>偽画像事案が急増し、対策を講じなければ自社が誤報の当事者になるリスク</li> <li>外部提供写真の真偽を確認する手段がない</li> <li>静止画だけでなく動画の偽造判定ニーズも増大</li> </ul>
拡散・影響に関する課題	<ul style="list-style-type: none"> <li>自治体の発信力ではデマの拡散力に太刀打ちできない</li> <li>問い合わせ殺到による業務圧迫、外国人排斥デマによる社会的混乱</li> </ul>	<ul style="list-style-type: none"> <li>複数部署に散発的に写真が提供され、単一部署での対応が困難</li> </ul>

## 5-2. 社会実装に向けた取組の個別詳細

### 2. 戦略的パートナーシップの構築

- 協業可能性がある事業者とパートナーシップ連携に向けた意見交換を実施した

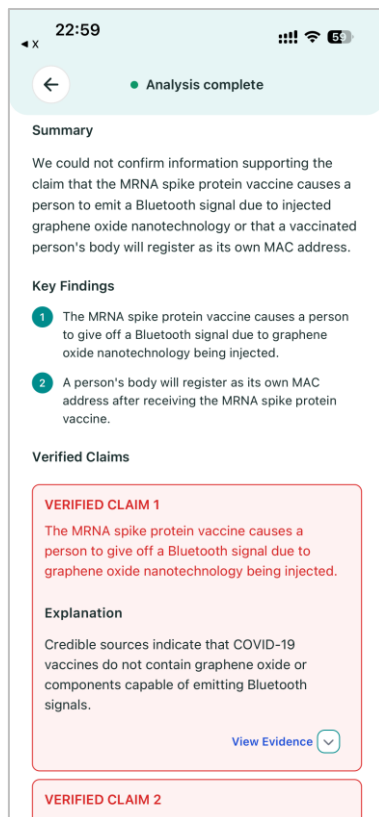
パートナー類型	ディスカッション目的	パートナー側の役割	弊社の役割
ソーシャルリスニング事業者	ソーシャルリスニングサービスと弊社AI分析機能の連携可能性を協議	SNS・ニュース・掲示板等からの大規模データ収集、一次的なトレンド・感情分析、APIによるデータ提供	取得データに対する独自のディープフェイク判定・フィルタリングを適用し、偽情報対策サービスとしてクライアントへ提供
偽情報コンサルティング事業者	OSINT・アナリストの知見と弊社AI技術の相互補完によるサービス可能性を協議	上流の要件定義・顧客対応、専門アナリストによる分析・運用、官公庁への販路提供	システム開発、AIモデルの提供・実装、技術的ソリューション提供

## 5-2. 社会実装に向けた取組の個別詳細

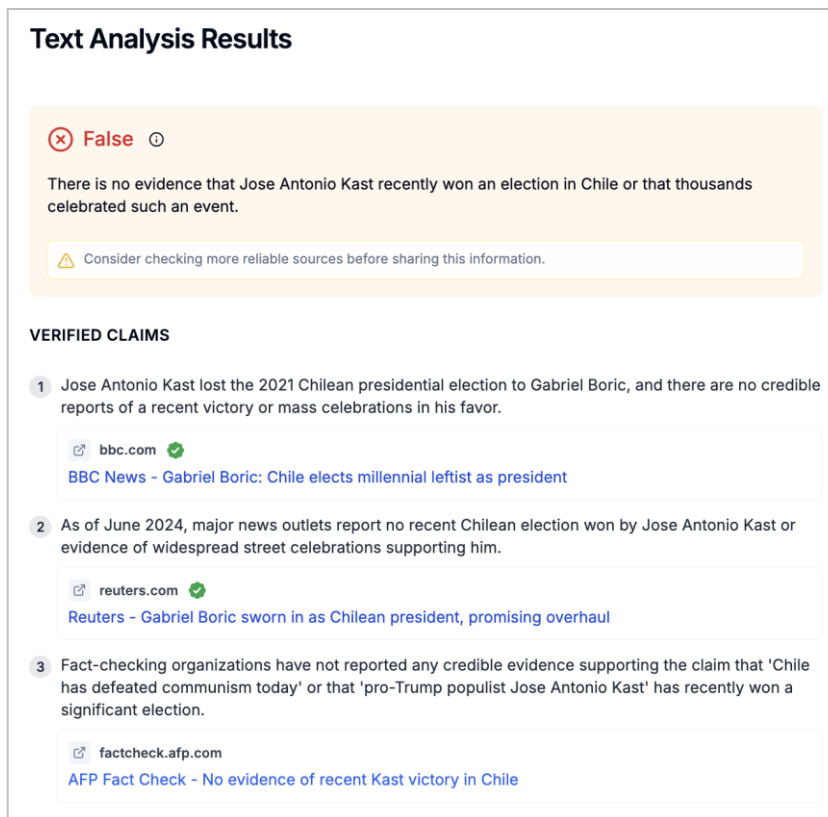
### 3. グローバル展開の検討 (1/2)

- 開発ツールは英語にも対応しており、今後他の言語にも対応することが可能である

#### SNSユーザー向けツール



#### 訂正主体向けツール



## 5-2. 社会実装に向けた取組の個別詳細

### 3. グローバル展開の検討 (2/2)

- 海外の主要スタートアップの動向を調査

企業名	拠点	コア技術・アプローチ	主なターゲット市場	特記事項・競争優位性
<b>Blackbird.AI</b>	米国	ナラティブ分析	大手企業、金融、政府	リスクの金銭的定量化、Gartner評価リーダー
<b>Reality Defender</b>	米国	ディープフェイク検知	銀行、政府、プラットフォーム	コールセンター向け音声検知、アンサンブルモデル
<b>Truepic</b>	米国	C2PA・来歴証明	保険、自動車、報道	撮影時の真正性証明、C2PA標準化の主導
<b>Cyabra</b>	イスラエル	ボット検知	官公庁、消費財ブランド	ボットネットワーク検知、NASDAQ上場予定
<b>ActiveFence (Alice)</b>	イスラエル	トラスト&セーフティ基盤	プラットフォーム、生成AI開発者	モデレーションのインフラ化、レッドチーミング機能
<b>Factiverse</b>	ノルウェー	自動ファクトチェック	メディア、防衛	リアルタイム動画検証、学術的バックボーン
<b>Gogolook</b>	台湾	詐欺検知・消費者保護	通信、金融、一般消費者	独自DBとAIの統合、ScamAdviser買収によるデータ網

# 目次

1. 開発・実証のサマリ
  1. 開発・実証のサマリ
2. 開発・実証の背景・目的
  1. 開発技術によりアプローチする課題
  2. 開発技術により目指す姿・ゴール
  3. 開発技術により対処可能なユースケース
3. 開発・実証における「対策技術の開発」
  1. 技術開発の全体像
  2. 技術開発の個別詳細
4. 開発・実証における「対策技術の有効性等に関する検証及び調査」
  1. 検証及び調査の全体像
  2. 検証及び調査の個別詳細
5. 開発・実証における「対策技術の社会実装に向けた取組」
  1. 社会実装に向けた取組の全体像
  2. 社会実装に向けた取組の個別詳細
6. 開発・実証における「普及啓発活動への協力」
  1. 普及啓発活動の全体像
  2. 普及啓発活動の個別詳細
7. 開発・実証の課題・展望
  1. 技術開発及び社会実装における課題・展望
8. 開発・実証の実施体制等
  1. 実施体制及び役割分担
  2. 全体スケジュール

## 6-1. 普及啓発活動の全体像

### 普及啓発活動に係る取組・成果の全体像

- 偽誤情報対策に関するメディア取材やイベントに積極的に登壇

日付	取材・イベント
2025年9月10日	NTTデータ「豊洲の港から」定例会において「Disinformation Security（偽・誤情報対策）」に登壇
2025年10月8日	日本経済新聞社主催「生成AIサミット」インパクトピッチのファイナリストに選出
2025年12月7日	テレビ東京「クリックニッポン」「偽・誤情報対策」に出演
2026年2月3日	毎日新聞にて掲載
2026年2月25日	共同通信より記事化

# 目次

1. 開発・実証のサマリ
  1. 開発・実証のサマリ
2. 開発・実証の背景・目的
  1. 開発技術によりアプローチする課題
  2. 開発技術により目指す姿・ゴール
  3. 開発技術により対処可能なユースケース
3. 開発・実証における「対策技術の開発」
  1. 技術開発の全体像
  2. 技術開発の個別詳細
4. 開発・実証における「対策技術の有効性等に関する検証及び調査」
  1. 検証及び調査の全体像
  2. 検証及び調査の個別詳細
5. 開発・実証における「対策技術の社会実装に向けた取組」
  1. 社会実装に向けた取組の全体像
  2. 社会実装に向けた取組の個別詳細
6. 開発・実証における「普及啓発活動への協力」
  1. 普及啓発活動の全体像
  2. 普及啓発活動の個別詳細
7. 開発・実証の課題・展望
  1. 技術開発及び社会実装における課題・展望
8. 開発・実証の実施体制等
  1. 実施体制及び役割分担
  2. 全体スケジュール

## 7-1. 技術開発及び社会実装における課題・展望

### 技術開発及び社会実装にあたっての今後の課題およびそれらを踏まえた今後の展望

- 本年度の開発・実証を通じて、技術開発面では「生成AIの急速な進化への追従」「分析コストの最適化」「検出対象メディアの拡大」「多言語対応の高度化」の4つが主要課題として明確化され、それぞれ具体的な対応策を策定した。

#### 技術開発面の課題と対応策

課題	背景・要因	対応策・展望
<b>生成AI技術の急速な進化への追従</b>	非連続的に進化する生成AIにより、新たなディープフェイクが次々と出現。既存モデルの精度劣化リスクが常に存在する。	本年度構築したMLOpsパイプライン（データ収集・監視→テスト→チューニングデータ生成）により、低コストかつ迅速なモデル改善サイクルを確立。次年度以降、このパイプラインの自動化率を向上させ、新種のディープフェイクへの対応リードタイムを短縮する。
<b>分析コストの最適化</b>	要素技術2（根拠情報検索）でのLLMコストがユーザー数増加に伴いコストが増大する。事業としての損益分岐点の達成が課題。	軽量モデルの活用やキャッシュ機構の導入、段階的な処理（簡易判定→詳細分析）により、1リクエストあたりの分析コストを最小化する。オンプレミス環境でのモデル運用も検討し、スケール時のコスト構造を最適化する。
<b>検出対象メディアの拡大</b>	現在の対応はテキスト・画像・動画・音声を中心であるが、ショート動画プラットフォームにおける偽・誤情報は増加傾向にあり、対応の遅れが懸念される。	ショート動画プラットフォームへの対応策を検討。動画視聴時などにリアルタイムの偽情報検出にも対応可能な技術基盤の構築を目指す。
<b>多言語対応の高度化</b>	英語対応は完了しているが、中国語・韓国語等からの越境的な偽情報キャンペーンへの対応ができない。	アプリの多言語対応を進めると同時に、各地域での配信・ユーザー獲得も検討する

## 7-1. 技術開発及び社会実装における課題・展望

### 技術開発及び社会実装にあたっての今後の課題およびそれらを踏まえた今後の展望

- 社会実装面では「SNSユーザーの獲得・定着」「訂正情報の流通促進」「訂正主体のツール活用促進」「法規制・倫理面への対応」の4つが主要課題であり、パートナーエコシステムの構築とUX設計の改善を軸に対応を進める。

#### 社会実装の課題と対応策

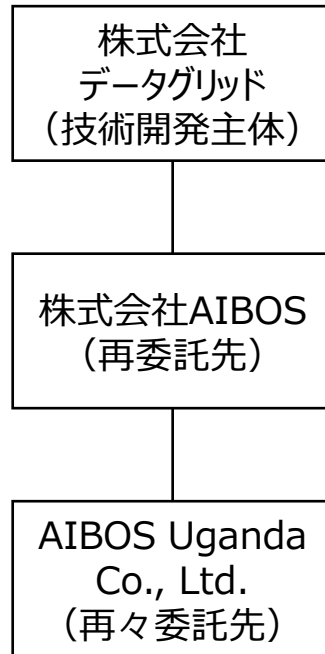
課題	背景・要因	対応策・展望
<b>SNSユーザーの獲得・定着</b>	パブリックベータ版テストを通じて一定の月間アクティブユーザーは確保できたが、継続的な利用定着には至っていないユーザー層も存在。ファクトチェック行為自体が日常的な習慣として定着していないことが根本的な課題。	検証結果の可視化やゲーミフィケーション要素の導入により利用動機を強化。SNSプラットフォーム上での自然な導線設計により、ユーザーの行動フローに溶け込むUX設計を実現する。
<b>訂正情報の流通促進</b>	ファクトチェック結果は作成されても、偽情報を信じたユーザーに十分にリーチできていない「訂正情報の非対称性」が構造的課題。	SNSユーザー向けツール「シラベル」を訂正情報の流通チャネルとして活用。訂正主体が発信する訂正情報を、ユーザーのファクトチェック結果として自動的に表示する仕組みを構築し、訂正情報の到達率を向上させる。
<b>訂正主体のツール活用促進</b>	ヒアリングを通じてニーズの強さは確認できたものの、既存業務フローへの組み込みや、組織内での導入決定プロセスにおける障壁が存在。	パートナー企業との協業を通じた間接的なアプローチにより、既存の顧客基盤を活用した導入促進を実施。導入事例の蓄積と共有によりリファレンスを構築する。
<b>法規制・倫理面への対応</b>	偽情報対策における表現の自由とのバランス、AIによる判定結果の説明責任、個人情報の取り扱い等、法規制・倫理面での整理が必要。	技術アドバイザーや法律専門家と連携し、ガイドラインの策定と運用ルールの整備を進める。ツールはあくまで「検証支援」であり、最終判断はユーザーに委ねるという設計思想を堅持する。

# 目次

1. 開発・実証のサマリ
  1. 開発・実証のサマリ
2. 開発・実証の背景・目的
  1. 開発技術によりアプローチする課題
  2. 開発技術により目指す姿・ゴール
  3. 開発技術により対処可能なユースケース
3. 開発・実証における「対策技術の開発」
  1. 技術開発の全体像
  2. 技術開発の個別詳細
4. 開発・実証における「対策技術の有効性等に関する検証及び調査」
  1. 検証及び調査の全体像
  2. 検証及び調査の個別詳細
5. 開発・実証における「対策技術の社会実装に向けた取組」
  1. 社会実装に向けた取組の全体像
  2. 社会実装に向けた取組の個別詳細
6. 開発・実証における「普及啓発活動への協力」
  1. 普及啓発活動の全体像
  2. 普及啓発活動の個別詳細
7. 開発・実証の課題・展望
  1. 技術開発及び社会実装における課題・展望
8. 開発・実証の実施体制等
  1. 実施体制及び役割分担
  2. 全体スケジュール

## 8-1. 実施体制及び役割分担

### 本事業の実施体制図



### 各団体の役割・業務範囲

- 株式会社データグリッド
  - 本事業全般の管理・統括業務、技術開発・システム設計・開発の実施
- 株式会社AIBOS
  - フロントエンドとバックエンド（サーバーサイド、データベース、API、インフラ）を実装を行うとともに及びAIBOS Uganda Co., Ltd.の品質管理・進捗管理を担当
- AIBOS Uganda Co., Ltd.
  - 株式会社AIBOSの品質管理・進捗管理体制のもと、フロントエンド開発、バックエンド開発、API開発、データベース設計を担当

## 8-2. 全体スケジュール

主な実施事項	令和7年						令和8年	
	8月	9月	10月	11月	12月	1月	2月	3月
(1)対策技術の開発								
要素技術1：ディープフェイク検出技術	データセット基盤構築		ディープフェイク判定基盤構築			性能評価及び改善		
要素技術2：根拠検索技術	設計	開発			性能評価及び改善			
要素技術3：文脈検証技術	設計	開発			性能評価及び改善			
要素技術4：情報接種・行動分析技術	設計	開発			性能評価及び改善			
要素技術5：リスクアセスメント・拡散予測技術	設計		開発			性能評価及び改善		
(2)対策技術の有効性等に関する検証及び調査								
SNSユーザー向けツールの設計・開発・運用	設計	詳細設計・開発			運用・改善			
検証1：クローズドアルファ版テスト					0ターム	1stターム	2ndターム	3rdターム
検証2：パブリックベータ版テスト							運用・改善	
訂正主体向けツールの設計・開発	設計	詳細設計・開発			運用・改善			
検証3：テスト検証							テスト検証	
(3)対策技術の社会実装に向けた取組								
1.ニーズの深掘りと事業計画の精緻化	ポテンシャルカスタマーとの意見交換によるニーズの深掘り				事業計画の精緻化作業			
2.戦略的パートナーシップの構築	パートナー候補先の探索、意見交換による連携案の具体化							
3.グローバル展開の検討					海外動向の調査・分析			
(4)成果報告書。社会実装実施計画書の作成					報告書の作成			
(5)普及啓発活動への協力					メディア対応、イベント登壇などの普及啓蒙活動を実施			