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

多元統合型偽・誤情報検出技術の開発・実証

成果報告書 概要版

2026/3/19

技12_サン電子株式会社

目次

目次

1-1. 開発技術によりアプローチする課題・目指す姿

ツール全体および新規開発要素の概要図

真偽情報判定基盤の新規開発
― 単独判定から統合判定へ ―

生成AI時代、単独ツールの限界を克服し、高精度・説明可能な統合判定基盤を構築して実施

生成AI時代の課題

- 真偽判定困難な誤情報、偽情報が拡散しやすい
- 真偽判定ツール間で結果不一致
- 判定根拠不透明で共通ルールの欠如

ツールA・B・C

ツールA	ツールB	ツールC
根拠簡略	中間曖昧	詳細不足
誤差変動	柔軟不足	傾向検知
定量評価	分類特化	媒体限定
数値出力	判定明快	詳細不足

真偽情報判定基盤の新規開発

判定結果

ツールA	ツールB	ツールC
Fake0.8	0.97	Suspicious

正規化 ※判定形式の統一化
異なる出力仕様を統一

ツールA	ツールB	ツールC
0.9	0.97	0.75

統合ロジック
ツールの特徴を活かした重み付け統合
Weighted統合方式
w1×0.9+w2×0.97+w3×0.75

目指す姿

- 単独判定から統合判定へ
- 説明可能かつ再現性のある社会実装基盤へ

統合判定結果

- 判定の高精度化
- 判定の安定化
- 履歴管理
- 出力形式の統一化
- 判定根拠の可視化

統合判定結果

- 統合スコア
- 信頼度
- 根拠内訳

単独の真偽判定ツールを統合することで、安定的な高精度判定と根拠の可視化を実現

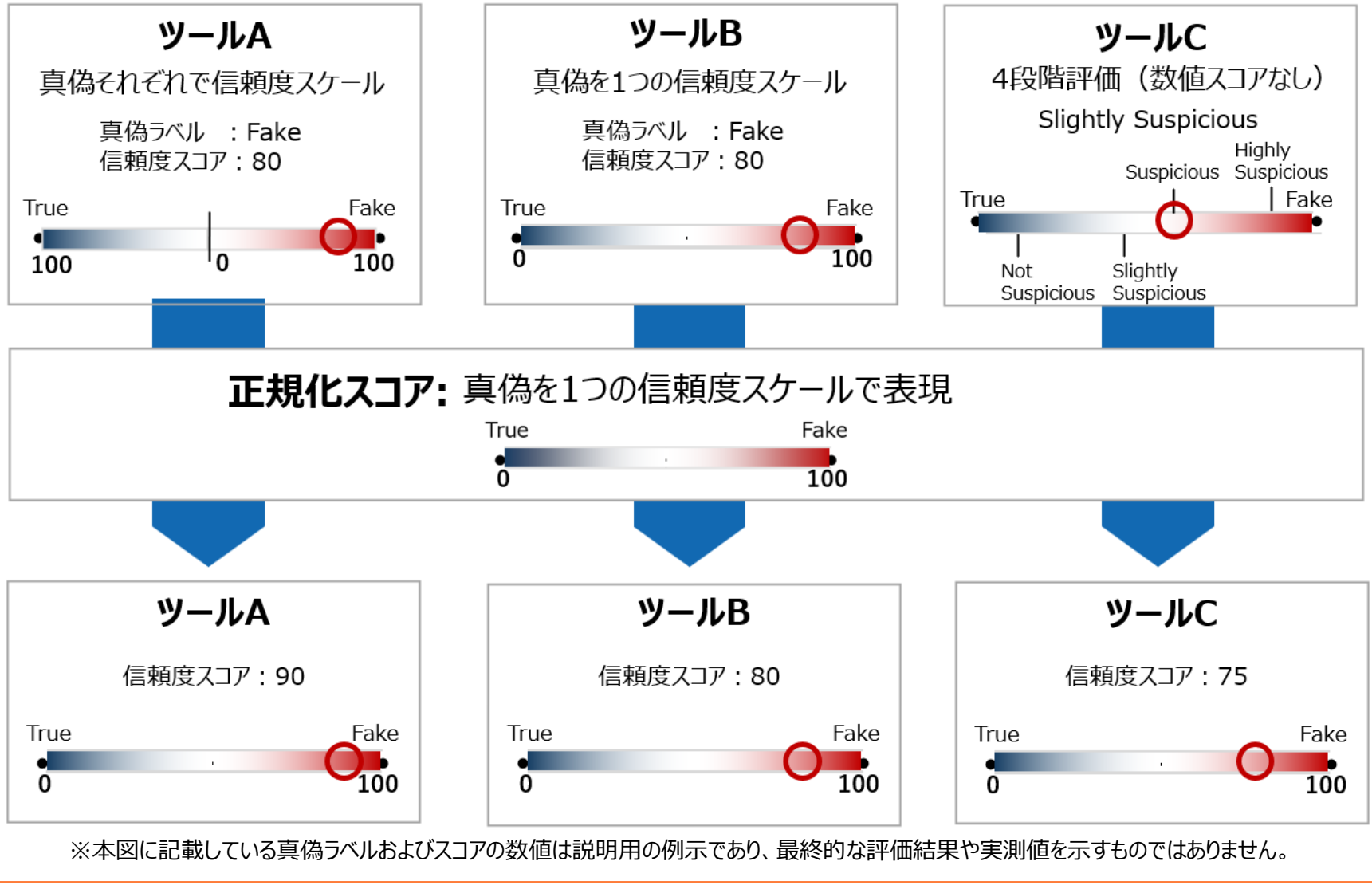
1-1. 開発技術によりアプローチする課題・目指す姿

各ツールの特徴

項目	ツールA	ツールB	ツールC
主な対象	画像・動画・音声対応	画像・動画・音声対応	画像・動画・SNS
出力形式	真偽ラベル 信頼度スコア 判定理由	真偽ラベル 信頼度スコア 悪意の可能性 スコア判定理由	4段階評価
強み	最新生成モデルの検出	文脈分析 URL参照	ボット・偽アカウント分析
分析観点	動画（フレーム解析＋音声同期） 音声（スペクトル・時間構造解析） 画像（ピクセル構造・周波数解析） ＝＞複数生成技術観点での解析	複数生成技術観点での解析 （検出モデルはブラックボックス） 文脈分析 公式サイトとの照合	拡散構造・ネットワーク分析
想定用途	真偽判定補助	改ざん検出	世論操作リスク検知

1-1. 開発技術によりアプローチする課題・目指す姿

異なる出力仕様を共通形式化



※本図に記載している真偽ラベルおよびスコアの数値は説明用の例示であり、最終的な評価結果や実測値を示すものではありません。

1-1. 開発技術によりアプローチする課題・目指す姿

開発技術によりアプローチする課題

- 近年、AIの進展により、画像・動画・音声・文章を容易に生成・加工できる環境が広がり、真偽が判別しにくい誤情報や偽情報の拡散リスクが高まっている。これらは世論誘導や信用失墜、社会的混乱を招くおそれがある。
- 現在は単一の判定ツールによる対応が中心だが、基準や得意分野の違いにより結果が一致しない場合があり、判定根拠の分かりにくさも課題である。
- 実務では判断基準が統一されておらず、属人的になりやすい。特に公共分野では再現性と説明責任が求められるため、複数技術を統合し、根拠を示せる仕組みの整備が必要である。

生成AIの普及による情報環境の変化と真偽判断の課題

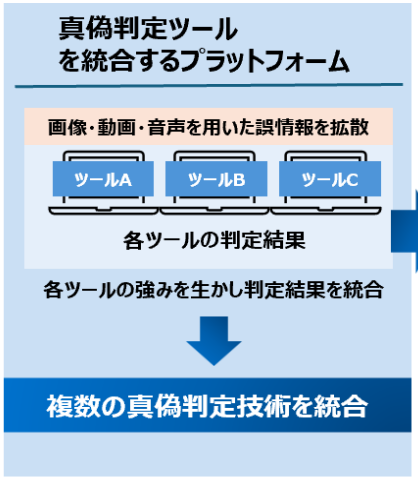
AIにより情報生成が容易化	誤情報・偽情報が拡散しやすい環境	公共性の高い分野での判断負荷
◆ 画像・動画・音声・文章を誰でも簡単に作成可能	◆ 事実と見分けがつかない情報が増加	◆ 行政・報道・企業では迅速かつ正確な判断が必要
◆ 専門知識や特別な設備が不要に	◆ 世論誘導・信用失墜・社会的混乱を招くおそれ	◆ 誤判断は対応遅れや信頼低下につながる
◆ 一般向けツールで偽の映像・音声も生成可能	◆ 悪意がなくても再拡散されるケースが増加	◆ 限られた時間で根拠ある判断が求められる

単一の真偽判定ツールでは様々なメディアに安定した精度を維持することが難しく、判断根拠の説明性を欠けることが多い

上記課題を踏まえ目指す姿・ゴール

- 異なる手法の複数ツールを統合し、コンテンツ横断で安定した高精度判定を行う「標準的な偽・誤情報判定基盤」の確立を目指す。各ツールの強みを活かし弱点を補完する統合ロジックを構築し、単独製品では得られない高信頼・高再現性の判定を実現する。
- さらに、判定根拠と評価指標を可視化して透明性を高め、行政・報道・企業が説明責任を果たしやすい基盤を整備する。ユースケース別に最適な重みや閾値を自動導出し、偽・誤情報対策の高度化・標準化・省力化に貢献する社会実装を最終目標とする。

複数技術を統合した行政機関向けの情報真偽判定システム



本実証技術による真偽判定におけるメリット

高度化

単一技術では見抜けない複雑な偽誤情報にも対応できる精度と適応力

標準化

組織を超えて共通の評価指標・判定プロセスを確立し、再現性と説明可能性を担保

省力化

経験に依存せずとも一定水準の判断が可能とし、人的負荷と判断時間を削減

1-2. 技術開発の取組・成果

検証用データの収集

- 取組内容
  - 本開発では、ディープフェイクに対する統合判定技術の有効性検証を目的に、検証用データを整備した。商用利用可能な公開データセットおよび生成AIで作成した真正／偽コンテンツを体系的に収集した。
  - 解像度・形式・圧縮方式・生成手法の違いを考慮し、特定条件に依存しない構成とした。
  - さらに、判別が容易なものから高度に精巧なものまで難易度別に整理し、実運用を想定した評価が可能なデータセットを構築した。
  - また、シーン全体を生成する汎用生成AIコンテンツも含め、従来手法では検出困難なケースにも対応した。ユースケース別に整理し、実社会を想定した検証を行った。

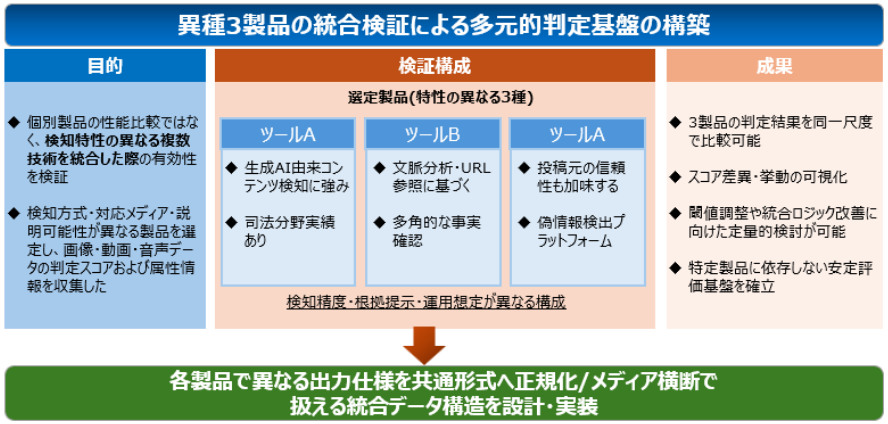
- 成果
  - 条件変化下でも統合判定技術の安定性を確認できる評価基盤を構築した。
  - 容易なケースから困難なケースまで客観的評価が可能となり、社会実装に向けた技術的課題と適用範囲を整理する知見を得た。

1-2. 技術開発の取組・成果

3製品の出力データの収集・正規化処理の実施

- **取組内容**
  - 本実証は、個別製品の性能評価ではなく、特性や出力形式の異なる複数技術を統合した際の有効性を検証することを目的とした。
  - 検知方式・対応コンテンツ・説明可能性が異なる3製品を選定し、画像・動画・音声データの判定スコアおよび属性情報を収集した。
  - ツールA：生成AI由来の偽造検知に強みを持つ真偽判定ツール
  - ツールB：文脈分析・URL参照に基づく多角的な事実確認ツール
  - ツールC：投稿元の信頼性も加味する偽情報検出プラットフォーム
  - 各製品で異なる出力仕様を共通形式へ正規化し、コンテンツ横断で扱える統合データ構造を設計・実装した。

- **成果**
  - 3製品の判定結果を同一尺度で比較可能とするため正規化処理を実装し、統合評価のためのデータ基盤を構築。
  - スコアや判定挙動の差異を可視化し、閾値調整や統合ロジック改善に向けた定量的検討が可能となった。
  - 実証期間を通じて、特定製品の出力形式に依存しない安定的な処理・評価環境を確立した。



1-2. 技術開発の取組・成果

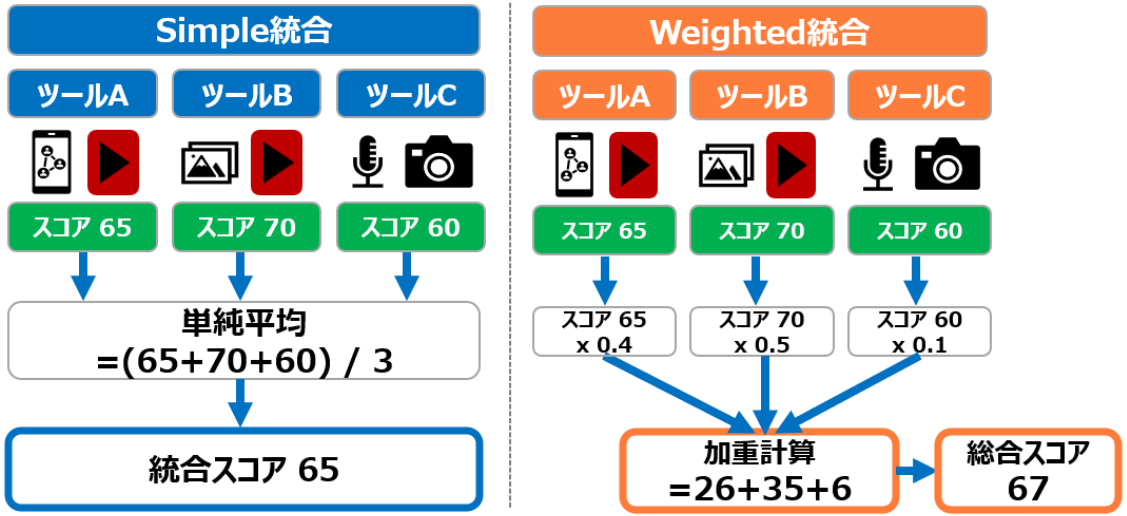
統合ロジックの設計・実装

取組内容

- 3製品のスコアを利用した2種類の統合方式（Simple、Weighted）を実装
- Simple統合は複数の真偽判定ツールの判定結果を同等に扱い、判定結果を単純平均する統合方式
- Weighted統合は検知特性や精度傾向を踏まえて、複数の真偽判定ツールの判定結果の重みづけを調整する統合方式
- 設計プロセスにおいて、コンテンツ特性（画像／動画／音声）別の重みや特性差も検討

成果

- 単一ツールによる判定と比較して、より高い精度を示す統合スコアを安定して算出可能
- コンテンツごとのツール特性を考慮した重み付けを算出し、コンテンツに応じた判定の最適化を実現
- ツールごとの強み・弱みが統合ロジックに反映され、過検知や見逃しの低減を実現



※統合スコアは各ツール出力を平均した総合評価値であり、精度(accuracy)を示すものではありません。

1-2. 技術開発の個別詳細

総括KPI（技術開発の達成度評価）

- 本開発・実証期間では、統合判定技術の社会実装可能性を総合的に評価する観点から、3つのKPIとして設定した。これは、技術の有効性を単一の性能指標のみで評価するのではなく、実運用を想定した適用範囲、判定性能、および社会実装に必要な基盤整備の3側面から総合的に評価することを目的としたものである。
- KPI①については主要ユースケース対応率の目標60%に対し67%を達成し、想定したユースケース群において統合判定技術が適用可能であることを確認した。
- KPI②については、画像を中心に精度向上および誤検知率低減の効果が確認され、画像においては精度が7.0%向上し、動画においては精度が7.5%向上した。また、誤検知率は最大で約50%の低減となるなど、統合判定による性能改善の有効性が確認された。
- KPI③については、判定結果正規化機能、統合判定ロジック、統合効果検証手法の確立、および社会実装実施計画の策定を計画どおり完了し、社会実装に向けた技術基盤を整備した。
- 以上より、本年度に設定したKPIは概ね達成され、統合判定技術が信頼性判断支援のための実用的な基盤として機能する可能性が示された。今後は、動画を含む多様なコンテンツへの適用性向上、評価データの拡充、ならびにユースケースごとの最適な重み付け・閾値設計の高度化を進めることで、適用領域の拡大と社会実装の具体化を図る予定である。

KPI項目	評価指標	本年度目標	達成度	評価
KPI① 信頼性判断支援ユースケースのカバレッジ	主要ユースケース対応率	60%	67%相当を達成	◎
KPI② 統合判定による精度・安定性向上	精度向上／誤検知率低減	精度 +5～+15% 誤検知率▲15%	精度 最大+7.5% 誤検知率 最大▲50%	◎
KPI③ 社会実装基盤の技術的完成度	基盤機能の設計・実装完了	基盤整備完了	計画通り完了	○

凡例
◎：目標達成（計画以上または想定通りの成果を確認）
○：概ね達成（社会実装に向けた基盤を確立）
△：一部未達（次年度以降に継続対応）

目次

2-1. 社会実装に係る取組・成果

社会実装に係る取組・成果

取組と成果の全体像

アクション	結果 / 気づき
公共・企業へのヒアリング調査	「判断支援」への課題や期待効果の確認
持続可能なビジネスモデルの構築検討	実用サービス展開の妥当性の確認
国内展開戦略・市場性の基本評価	他団体との連携と展開ロードマップの想定
多元統合型判定による検知精度の実証	高度偽造コンテンツに対する有効性
事業化における経済価値の特定	監視工数削減等の具体的ROIの可視化

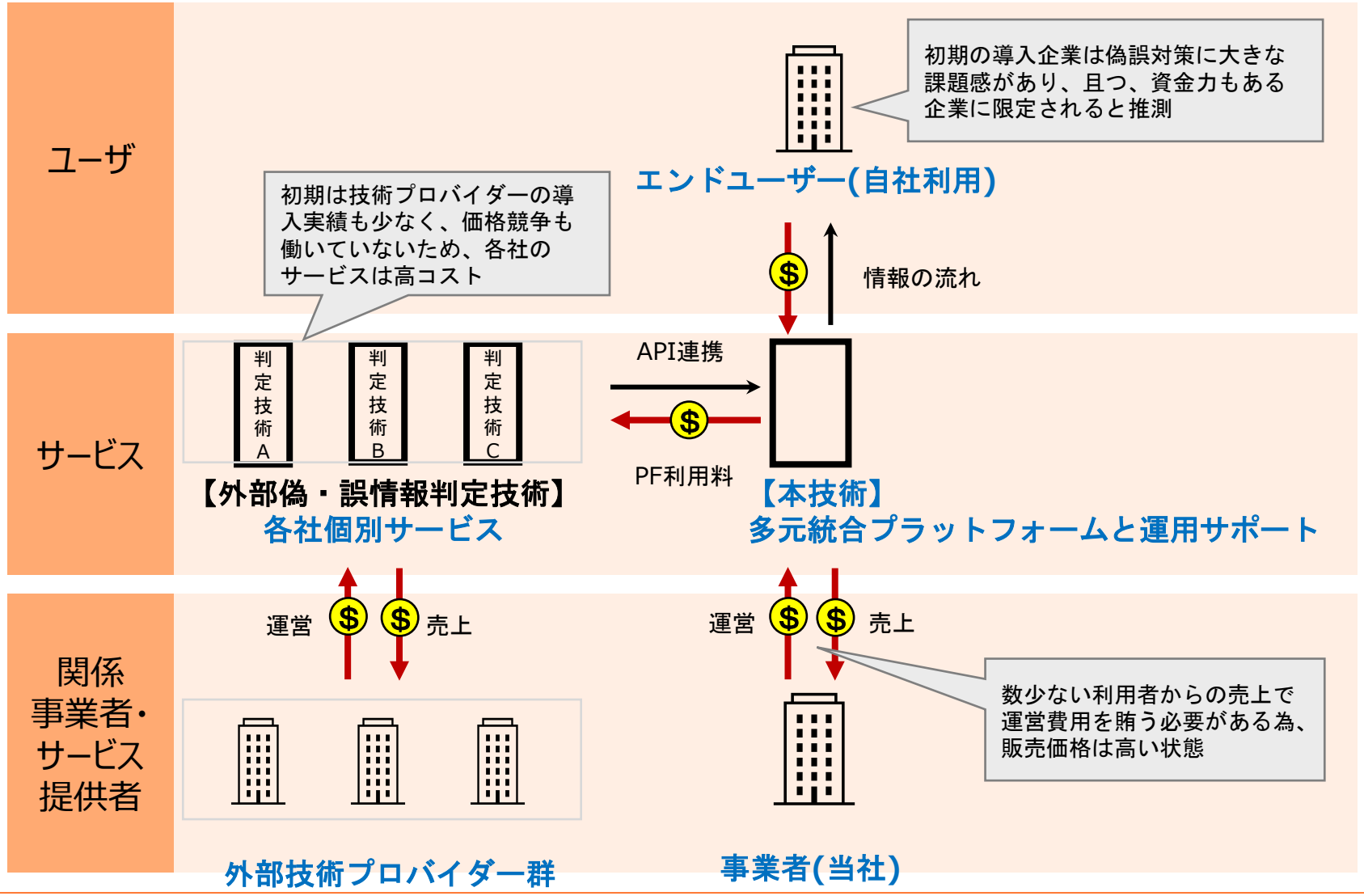
2-1. 社会実装に係る取組・成果

社会実装に係る取組・成果

- 公共機関や企業へのヒアリングを通じ、公共への適応性や活用可能性を把握するとともに懸念点解消に向けて下記の基本課題の調査に取り組んだ。
  - 現行の対策プロセスや課題把握
  - 導入障壁や懸念事項(コスト/運用負荷/法的懸念等)の明確化
  - 最初に実装すべきユースケースや期待効果の確認
- 持続可能なビジネスモデルを構築するために、実用的なサービスの展開を見据えた取組を行った。
  - 仮説設定した本技術の導入フェーズ別のビジネスモデルの妥当性検討
  - 顧客視点での価格受容性や社会への普及等の見極め
- 国内展開市場を狙っていくための取組として、展開戦略の基本検討・市場性/事業性の基本評価を行った。
  - 社会実装の加速を目的とした他事業者とのパートナーシップや協業の検討
    特に、他団体、企業との連携としてコンソーシアムの参加の必要性を認識した。
  - 自社視点での持続可能な事業として社会への普及の時間軸と収益性評価(キャズムの克服検討)
- 検知精度の有意な向上
  - 統合判定技術の導入により、根拠を示した判定支援の必要性を確認した。特に生成AIで作成したディープフェイク等の高度な偽造コンテンツに対する頑健性が向上した。
- 事業化における具体的な経済価値の特定
  - ターゲット層において、現状年間3,000万円程度発生している監視・法務コストを本サービスで代替・効率化できるという具体的なROIを算出した。
  - 「3,000万円規模の投資判断」が可能な市場ニーズが存在することを特定し、売上具体化に向けた確度の高いビジネスモデルを構築できた。

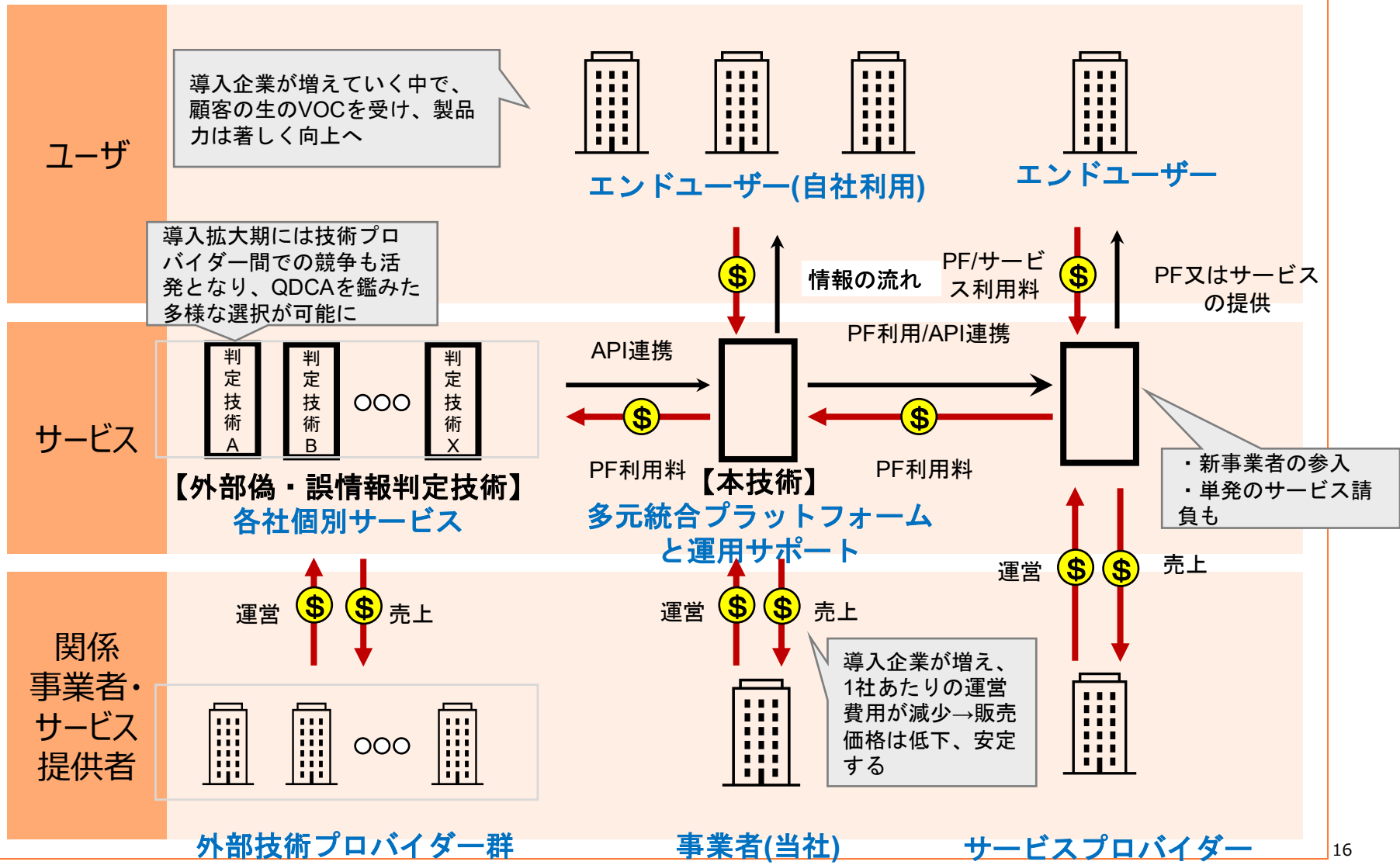
2-2. 社会実装時のビジネスモデル

社会実装時のビジネスモデル　①初期導入フェーズ



2-2. 社会実装時のビジネスモデル

社会実装時のビジネスモデル　②導入拡大フェーズ



2-2. 社会実装時のビジネスモデル等

概要				
ユースケース	具体的な対象	ペインポイント（課題）	本技術による解決策	ユーザ視点での効用
エンターテインメント/IPホルダー	・芸能プロダクション、インフルエンサー事務所、コンテンツ制作会社など	・膨大な監視・対処コスト ・ブランド価値毀損と機会損失 ・検知後の対処フローの断絶	・多元統合型判定技術による検知と、ワンクリックでの他部署連携用レポート生成。	・監視・初動対応コストの劇的な削減（ROIの向上） ・即時否定声明が可能に。
官公庁/地方自治体	・中央省庁、地方自治体の広報、防災部門、警察・法執行機関など	・災害・有事のデマ拡散 ・公的な信頼性の低下 ・説明責任の負担	・多元統合型判定による信頼度スコアリングと、判定根拠の可視化。	・行政判断の正確性と迅速性の向上。 ・説明責任を円滑に果たせる。
報道機関/メディア	・新聞社、テレビ局、ネットニュース運営会社など	・真偽判定の工数限界 ・高度な偽造への対応 ・誤報リスクへの懸念	・複数の検知結果の統合による誤判定リスクの低減と、判定根拠の可視化。	・真偽判定業務の効率化と報道精度の向上。 ・迅速な放映判断が可能。
SNS/オンラインプラットフォーム運営者	・SNS運営会社、動画配信プラットフォームなど	・ソーシャルリスニングの限界 ・自社開発のコストと技術の風化 ・プラットフォームの信頼性低下	・統合判定ロジックによる大量スクリーニングと、API連携プランの提供。	・ソーシャルリスニング業務の劇的な省力化と品質向上。 ・技術の進化に柔軟に対応。

2-2. 社会実装時のビジネスモデル等

エンターテインメント・IPホルダー（タレントマネジメント等）/官公庁・地方自治体

具体的な対象	芸能プロダクション、インフルエンサー事務所、コンテンツ制作会社など。
ペインポイント	◆ 膨大な監視・対処コスト SNS上のなりすましやディープフェイクを人力で監視しており、年間2,000万～3,000万円相当の人件費・法務コストが発生している。 ◆ ブランド価値毀損と機会損失 フェイク動画の拡散によるイメージ悪化、スポンサー契約への致命的な悪影響。 ◆ 対処フローの断絶 検知後の「削除申請」や「法務部署・警察への連携」の初動が遅れコストも発生する。
本技術による解決と効用	**【解決策】**多元統合型判定技術による検知と、他部署連携用レポート生成。 **【効用】**監視・初動対応コストの劇的な削減（ROIの向上）。客観的スコアを「心理的な許可証」として、担当者が自信を持って即時否定声明を出せるようになる。

具体的な対象	官公庁、地方自治体の広報・防災部門、警察・法執行機関など。
ペインポイント	◆ 災害・有事のデマ拡散 避難情報や天候に関する偽・誤情報が発生した際、正確な情報を届けるための真偽判定リソースが不足している。 ◆ 公的な信頼性の維持 偽・誤情報を放置することで、行政判断の遅れや市民からの不信感を招くリスク。 ◆ 説明責任の負担 なぜその情報を偽と判断したのか、第三者に合理的に説明するプロセスが未整備。
本技術による解決と効用	**【解決策】**多元統合型判定による信頼度スコアリングと、判定根拠の可視化。 **【効用】**行政判断の正確性と迅速性の向上。根拠を透明に提示できるため、住民や報道に対する説明責任を円滑に果たせる。

2-2. 社会実装時のビジネスモデル等

報道機関・メディア/SNS・オンラインプラットフォーム運営者

具体的な対象	新聞社、テレビ局、ネットニュース運営会社など。
ペインポイント	◆ 真偽判定の工数限界 SNS由来の素材確認に時間を要し、報道の即時性が損なわれる。 ◆ 高度な偽造への対応 専門知識がなければ判別不可能なディープフェイク（AI音声・動画）の流入。 ◆ 誤報リスクへの懸念 現状は「疑わしくは放映せず」の姿勢だが、その間に他媒体に先を越される機会損失。
本技術による解決と効用	**【解決策】**複数の検知結果の統合による誤判定リスクの低減と、判定根拠の可視化。 **【効用】**真偽判定業務の効率化と報道精度の向上。 編集会議等で判定根拠を共有でき、客観的エビデンスに基づく迅速な放映判断が可能となる。

具体的な対象	SNS運営会社、動画配信プラットフォームなど。
ペインポイント	◆ ソーシャルリスニングの限界 日々大量に投稿されるコンテンツを人手で全て確認することは不可能。 ◆ 自社開発のコストと技術の風化 高精度な判定エンジンを自社開発し続けるためのコストや人材が不足している。 ◆ プラットフォームの信頼性低下 フェイクやなりすましが蔓延することで、一般ユーザが離脱するリスク。
本技術による解決と効用	**【解決策】**統合判定ロジックによる大量スクリーニングと、API連携プランの提供。 **【効用】**ソーシャルリスニング業務の劇的な省力化と品質向上。 特定のツールに依存しない統合プラットフォームのため、 技術の進化（いたちごっこ）に柔軟に対応し続けられる。

2-3. 技術開発及び社会実装にあたっての課題・展望

技術開発及び社会実装にあたっての今後の課題

- 技術的限界とユーザー期待値の乖離
  - 本技術の事業化において、解決すべき課題は、ツールに対する過度な期待と実力値のギャップである。
  - 『これは偽物である』という断定的な判定を求めるユーザーの期待に対して、実際は「ツールAは80%偽物、ツールBは30%偽物、なぜならば～」という「判定支援」を行うことが市場に受容されるかが課題である。

- キャズム※の克服と社会接続への障壁
  - 持続可能なビジネスとして確立するためには、単なる「技術の提供」を超えた市場への適応が求められると予想。
  - 適応には1.ワークフローへの適応、2.心理的・法的適応、3.合理的コスト適応があると考えられる。

※キャズム：目新しさが評価される初期市場から、実用性を重視するメインストリームに移行する際に直面する「深い溝」のこと。

上記課題を踏まえた今後の展望

- 説明責任を果たす「判断支援」への高度化
  - 技術的課題に対し、AIに全権を委ねるのではなく、人間の意思決定を支える「高度な判断支援システム」への進化を目指す。

- 他事業者との協働による社会実装の加速
  - 運用的・市場的課題に対し、デザインパートナー※を巻き込み実効性の高い社会実装を推進する。

※デザインパートナーとは単なるテストユーザーではなく、課題発見から解決策の検証までを共創する存在として定義（例として保険会社、総合エンターテインメント企業など）

2-4. 事業の拡大に向けた中長期的な計画

事業の拡大に向けた中長期的な計画

- 本技術の事業化に向けた技術精度の継続的な磨きこみは事業者として行っていく一方、事業化の中長期的な課題としては、以下を想定している。
- 想定したターゲット顧客層に対して、本技術が導入拡大していく過程で「Nice to have」から「Must have」へのキャズムを乗り越える具体的施策の検討、実行（啓発活動含む）。
- 当社技術が目指す「信頼できる第三者の判定結果」の具体的な基準設計の検討、実行（例：規制強化と表現の自由のバランス等）。

フェーズ1（2026年度）ビジネスモデルの確立と初期実装

- 統合判定ロジックの開発・実証
  - ✓ 技術プロバイダーの単一技術を統合して、より精度を向上した判定結果を出力する技術を開発・実証する。
- 高ニーズ層への初期導入
  - ✓ 知財保護やブランド価値毀損に強い危機感を持つ大手エンタメ企業やIPホルダーを主ターゲットとし、占有型SaaSとしての導入を開始する。
- 収益構造の確立
  - ✓ 偽・誤情報対策を「経営リスク対策」として位置づけ、取締役決裁ラインの危機管理予算からの拠出を前提としたビジネスモデルを確立する。

フェーズ2（2027年度）ターゲットの拡大とワークフローの統合

- 「検知から対処まで」の一気通貫支援
  - ✓ 判定結果に基づく「社内報告用レポート」や「法的通報資料」の自動生成機能を深化させ、ワークフローを効率化・自動化する。
- 公共・金融領域への水平展開
  - ✓ 導入実績（リファレンス）を積み上げ、地方自治体、警察機関、金融機関など、情報の信頼性を最重視するより広範な市場へ展開を拡大する。
- エンタメ特有コンテキストの精緻化
  - ✓ Vtuberや実在タレントを模した高度な偽造を、エンタメ特有の演出や文脈から切り離して正確に検知できる体制を構築する。

フェーズ3（2028年度）社会防衛インフラとしての標準化と普及

- 「Must have」インフラへの定着
  - ✓ キャズムを克服し、デジタル社会の健全な発展を支える不可欠な「情報の防波堤」としての地位を確立する。
- 第三者判定の標準化（デファクトスタンダード化）
  - ✓ 特定の製品に依存しない「集合知としての判定」を「信頼できる第三者の判定結果」として社会的に定着させる基準設計を完遂する。
- スケールメリットによるコスト低減
  - ✓ 導入企業の拡大により1社あたりの運営費用を抑え、販売価格の低下と安定化を実現し、中堅・中小企業まで含めた幅広い普及を図る。