メディアの環境・形態が生成 AI を用いた詐欺への騙されやすさに与える影響の検討

A Study of the Impacts of Media Environment and Media Type on Vulnerability to Generative AI-Enabled Fraud

安全 20-0165　外村　秀仁
Shuji TONOMURA
指導教員：河野　和宏

This study focuses on the influences of Media environments and types on AI-generated fraud. We set scenarios with formal and casual environments using four media types: text, audio, image, and video. We then assess the credibility of four contents: Human-crafted real content, AI-generated fake content, Human-crafted fake content, and AI-generated real content. As a result, we show that AI-generated fake text content is most credible in both environments. Moreover, our analysis reveals that in casual environments, people trust AI-generated fake audio and video content based on a humanistic approach and subjective evaluation.

Key Words: *generative AI, fraud, fake, media environment, media type*

1. はじめに

近年，SNS 上での詐欺や偽情報が多数横行し，問題視されている．この理由の一つに，生成 AI 技術の発展がある．生成 AI を使うことで，誰もが簡単に，架空とは思えない高精度な文章，音声，画像を生成可能になったためである．

これまでのところ，偽情報においては，文章よりも音声，音声よりも画像の方が信じられやすい傾向にあり，音声や映像は関心を呼びやすいことが報告されているが，詐欺への影響は未知な部分が多い．特に生成 AI が用いられた場合にどこまで騙されやすさに影響があるかは不明である．

本研究では，生成 AI を用いた詐欺を対象に，メディア形態（テキスト，音声，画像，動画）と偽装対象の業種に基づくメディア環境（ソーシャルメディア，金融機関・Webmail/SaaS）が騙されやすさに与える影響を調べる．

2. なりすまし詐欺と虚偽情報による詐欺を想定したシナリオの実験概要

まず偽装対象の業種傾向[1]に基づき，想定シナリオのメディア環境をフォーマル環境（金融機関・Webmail/SaaS）とカジュアル環境（ソーシャルメディア）の 2 つに分類した．フォーマル環境は，個人向け通知環境で「企業の信頼性を重視したコンテンツ内容」とし，銀行の投資信託をシナリオの題材に選んだ．一方，カジュアル環境は，不特定多数向けの投稿環境で「企業の魅力度を重視したコンテンツ内容」とし，画像・映像の編集ソフトをテーマにした．

次に，2 種類の実験シナリオを準備した，実験 1 では，なりすまし詐欺を想定し，提示したリアル（実際の正しい情報）とフェイクの 2 つのコンテンツのうち，どちらが本物かという真偽判断を求めた．一方，実験 2 では，虚偽情報による詐欺を想定し，人による架空の内容（人による偽情報）と AI による実在の内容という 2 つのコンテンツに対して，どちらがより信頼できるか判断してもらった．これらの実験では，真偽判断や信頼性判断の質問に加え，判断理由の自由記述と，提示したコンテンツに対する 8 つの信頼性評価に関する 5 段階評価項目（好感，説得力，専門性，誠実さ，善意，明確さ，興味，独創性）を設けた．

(1) 実験 1 におけるリアルとフェイクの作成方法

リアルは「筆者が作成したオリジナル」なものであり，画像のみパブリックドメインの素材」を利用した．一方，フェイクは「生成 AI サービス・AI 変換技術で作成したもの」である．テキストと画像はオリジナルから ChatGPT・Adobe Firefly が生成したフェイク文章・フェイク画像，音声は ElevenLabs で作成した筆者のクローンボイスを用いて別人の声をボイスチェンジしたフェイク音声，動画は AKOOL で別人の顔を筆者の顔と交換したフェイク動画である．なお，フェイク音声とフェイク動画で話す文章は ChatGPT が生成したフェイク文章である．

フォーマル環境のシナリオは，セミナー案内メール，商品説明の音声案内，本社の画像，商品の月次報告動画の構成である．カジュアル環境のシナリオは，SNS におけるキャンペーン広告，機能説明のデモ音声，商品の広告画像（**図 1** 参照），クリエーターのインタビュー動画とした．

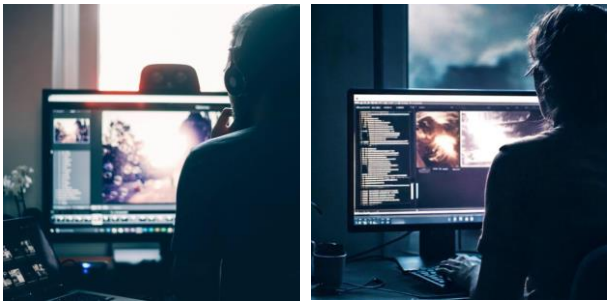

図 1 カジュアル環境のリアル画像（左）とフェイク画像（右）

(2) 実験 2 における人による架空の内容と AI による実在の内容の作成方法

人による架空の内容は「筆者が作成した実在しない内容を本物らしく説明した内容」である．一方，AI による実在の内容は ChatGPT により作成した「AI が実在の内容を説明した内容」である．テキストのみのシナリオ構成で，フォーマル環境では投資信託の商品説明，カジュアル環境では，編集ソフトの製品機能の説明という内容である．

3. 実験結果

2 つのメディア環境でそれぞれ，関西大学の学生 11 名，10 名から回答を得た．以下，各実験の結果を示す．

(1) 実験 1 の結果

表 1 の真偽判断の結果より，両環境でテキストは正解が約 20%と最も騙されやすく，画像は 63.7%以上の正解で見抜きやすいことがわかった．音声と動画はカジュアル環境で 40%～50%と正解が少なく，両環境で異なる傾向がある．

表 2 は，メディア環境ごとに，リアルとフェイクに対する評価項目を平均値の比較で分類した結果である．テキストは他のメディア形態より AI の説得力の影響が強く現れ，両環境でフェイクの方が機能性項目（好感，説得力，専門性，明確さ，興味，独創性）が高く評価される傾向にある．

一方，音声と動画はメディア環境の影響が生じていることがわかる．フォーマル環境では人間性項目（誠実さ，善意）と機能性項目が 2 軸に分かれる傾向があり，肯定的評価が約 20%と低い一方，カジュアル環境では全体的にリアルが 53.1%と高く評価され、フェイクにも 43.1%と高い評価であった．

(2) 実験 2 の結果

表 3 は実験 2 における信頼性判断の結果である．フォーマル環境では，人による架空の内容と AI による実在の内容が同程度の信頼性で，カジュアル環境では AI による実在の内容の方が信頼されやすい傾向にある．その一方，人による架空の内容に着目すると，フォーマル環境でより許容されることがわかった．

表 1 メディア環境別メディア形態と真偽判断のクロス集計表

メディア環境	メディア形態	正解	不正解	わからない
フォーマル環境	テキスト	27.3%	27.3%	45.5%
	音声	72.7%	18.2%	9.1%
	画像	63.6%	18.2%	18.2%
	動画	63.6%	0%	36.4%
カジュアル環境	テキスト	20%	60%	20%
	音声	50%	40%	10%
	画像	90%	0%	10%
	動画	40%	40%	20%

表 2 メディア環境別メディア形態ごとのリアルとフェイクに対する高評価項目（肯定的評価の割合の平均値）

メディア環境	メディア形態	リアル	フェイク
フォーマル環境	テキスト 画像		人間性項目 機能性項目
	音声 動画	人間性項目 (21.9%)	機能性項目 (28.2%)
カジュアル環境	テキスト	人間性項目	機能性項目
	音声 画像 動画	人間性項目 機能性項目 (53.1%)	 (43.1%)

表 3 メディア環境と信頼性判断のクロス集計表

メディア環境	人による架空の内容	AIによる架空の内容	わからない
フォーマル環境	40%	40%	20%
カジュアル環境	27.3%	45.5%	27.3%

4. 考察：まとめにかえて

テキストでは機能性（文章の精巧さ）が信頼性に影響する一方，音声，動画ではフォーマル環境で企業の信頼性に機能性が重視され，評価が厳格になり、カジュアル環境では企業の魅力度に人間性が重視され、評価が寛容になることがわかった．

真偽判断の理由から，フォーマル環境では「文章」や「動き」など客観的な基準，カジュアル環境では「印象」や「服装」など主観的な基準による評価方法の違いが，音声，動画は両環境で騙されやすさに影響したと推察する．

信頼性判断の理由では，フォーマル環境で説明のわかりやすさに関する記述が 30%あり，事実に基づかない内容でも平易な文章であれば信頼されることがわかった．

参考文献

[1] APWG：Phishing Attack Trends Report, https://apwg.org/trendsreports/（2025 年 1 月 11 日確認）．