

AIを用いたWEB動画の安全性を判定するシステムの開発

大阪教育大学附属高等学校池田校舎 2年生 岸本宗大, 沼田 蒼人, 多田 侑生, 田中創太, 増田 悠真

要旨 SNS上のAI動画による子どもへの悪影響が懸念される中、既存対策は不十分である。本研究は子どもを保護する制限基準の確立を目指し、暴力・フェイク・倫理・過刺激の観点から104項目の基準を作成・検証した。その結果、リスクの客観的評価の可能性が示された。今後は精度を高め、容易に活用できる評価システムの実用化を目指す。

1. 研究概要

研究背景

AI生成による偽広告やSNS上のAI動画が急増しており、特に子どもが真偽を判別できない現状がある。既存の動画サイトにおける制限では、**悪影響のある動画を排除しきれておらず**、プラットフォーム側からの新たなアプローチが必要であると考えた。

リサーチクエスト

新たなAI動画の評価方法を確立し、それを基に動画サイトで表示制限ルールを設けることで、子どもをはじめとする利用者への悪影響を防げるか。

仮説

現在の動画サイトはAI動画を適正に評価する基準が不十分なため、子どもが誤った情報や偏った価値観を信じてしまうリスクがある。新たな評価方法を基に表示制限や注意喚起を行うことで、この悪影響を軽減できる。

研究内容

本研究では、**AI動画の評価基準を作成して実際の動画を分析・評価**する。これにより、子どもが誤った知識や偏った価値観を受け入れるリスクを未然に防ぐ方法を検討する。

2. 先行研究

先行研究① AIによる声の模倣と権利侵害

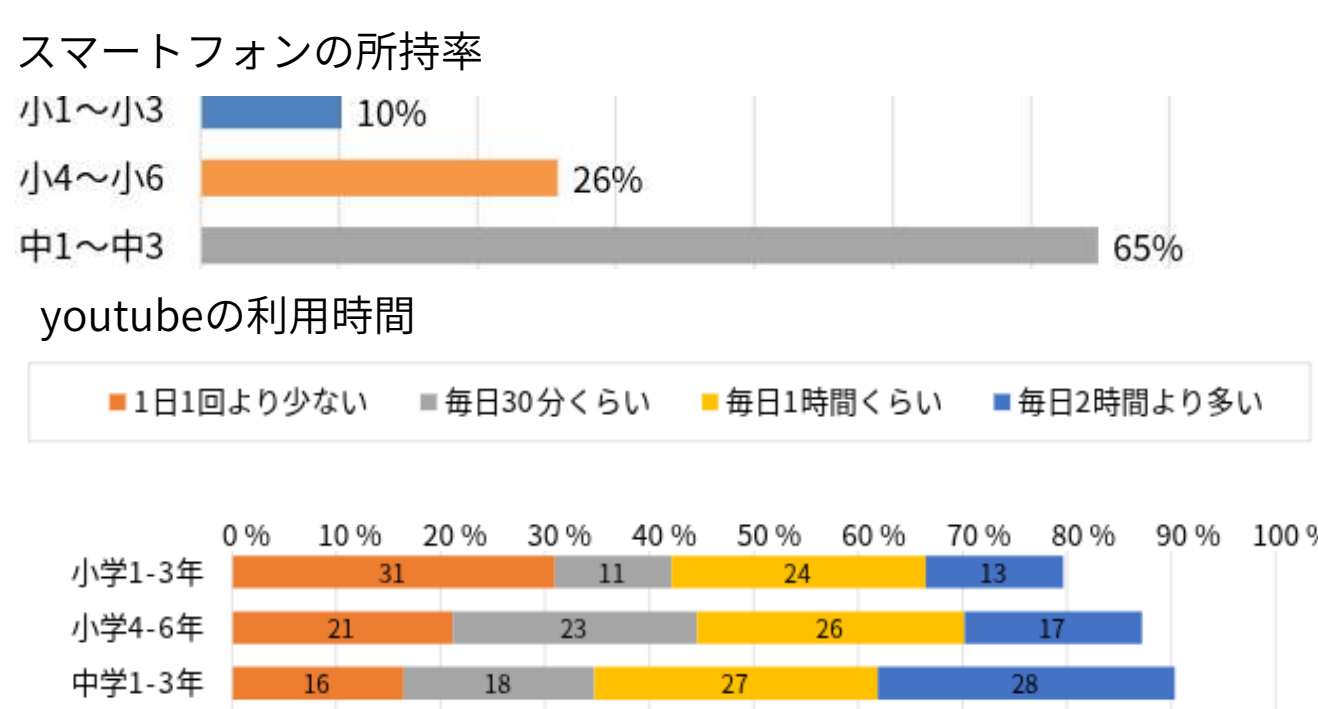
人気声優が無断でAI合成された声動画の削除を求め、TikTok運営を提訴した。急速に発展するAI音声技術に対し、許可なき無断利用を防ぐルール作りや運営側の対応が求められている。[1]

先行研究② 子どものインターネット・YouTube利用の実態

小学生の8割超がYouTubeを使い、中学生の約6割が自分用のスマホを持つなど動画視聴が日常化している。AI生成動画に触れる機会も増える中、安全な環境整備とメディアリテラシーの育成が急務である。[2]

先行研究③ AI動画が子どもに与える影響

AI動画は子どもが現実と区別しにくく誤情報を信じる危険があるとして、専門家がYouTubeに対策を打つように提言。技術向上に伴い、子どもを守る評価基準や規制の整備が急務である。[3]



参考資料②実際のAI動画の画像
(<https://www.youtube.com/watch?v=-AqndPZehHs>)

3. 活動内容

① 動画規制ガイドラインの作成

Googleの生成AI「Gemini」を活用し、ネット動画の安全性を評価する独自の違反判定基準を作成した。見落としがちな細かい表現まで計100項目以上をチェックし、違反を点数化することで、従来の動画サイトでは曖昧な判断が下されることの多い軽微な注意から公開停止まで**4段階で規制する仕組みを構築**した。

② 動画の自動チェック Python (Google Colaboratory) を用いて違反を自動検出するシステムを構築し、テスト用動画を読み込ませて**評価基準の実用性を検証**した。

③ 問題点の抽出とシステムの再設計 分析結果を多角的に検証・分類し、抽出された課題をもとに**システムを再設計**した。

④ 再実験と考察 改修後のシステムで再度チェックを行い、初期データと比較検証した。その結果、1回目の分析で課題として出た「動画の長さによる判定の不公平さの改善」や「一発で公開停止となるルールの正常動作」を確認できた。

「プログラム作成の様子」

AIに指示をして、mp4形式のデータを分析させるプログラム。

```
video_path = "/content/18.mp4"

print("1/4: 動画をAIのサーバーにアップロード中...")
video_file = client.files.upload(file=video_path)
print("2/4: アップロード完了!")

# 動画が利用可能 (ACTIVE) になるまで待つ処理
print("3/4: AIが動画を処理しています。準備ができるまでお待ちください...")
while True:
    check_file = client.files.get(name=video_file.name)
    if check_file.state.name == "ACTIVE":
        print("4/4: 動画の準備が完了しました!")
        break
    elif check_file.state.name == "FAILED":
        raise Exception("動画の処理に失敗しました。ファイルが破損している可能性があります。")
    time.sleep(2)

# 3. 104項目すべてを網羅した評価基準ルールブック (プロンプト)
prompt = """
あなたは動画の危険性を評価する専門AIです。

添付された動画（映像、音声、文脈、テロップ等）を詳細に分析し、以下の「評価基準」「重大な違反」に

【重大な違反】自殺・自傷行為の直接的・間接的な描写

殺人や致死行為にいたるプロセスのリアルな描写

テロ行為、監禁、誘拐、人質拘束の描写
```

5. 考察と今後の展望

AI動画の適切な評価にはリスクと関連する項目の厳選と**AI・人の併用**が必要であり、危険シーンの「秒数」に応じた評価体制の確立が不可欠であると分かった。しかし、今回の動画の評価に用いた基準は、法律に違反している内容ではないが、視聴する年齢によっては悪影響となる、「グレーゾーン」の基準の作成根拠が曖昧だった。そのため、表現の自由の保障と動画視聴者の保護のバランスがまだまだ粗く、実用化にはそのハードルを乗り越えなければいけないと感じた。今後は分析データを増やして**基準の信頼性を高める**とともに、URL入力で**危険性を自動判定**するWebサイトを開発し、不適切な動画から子どもを守る環境づくりに貢献したい。

参考文献

- [1] 日本経済新聞社. “AIで声模倣” 人気声優がTikTok運営会社提訴 動画削除請求” 日経電子版,2026-05-23,<https://www.nikkei.com/article/DGXZQOUD231K30T20C26A5000000/> (参照:2026-07-07).
- [2] モバイル社会研究所. “小中学生のYouTube利用率は8割超、スマホ所有率の上昇にともない増加傾向”. NTTドコモ モバイル社会研究所,2021-11-17, <https://www.moba-ken.jp/project/children/kodomo20211117.html> (参照 2026-07-07).
- [3]日本経済新聞社. “AI使ったYouTube動画「子供は現実と区別つかず」 米専門家らが抗議” 日経電子版,2026-04-02, <https://www.nikkei.com/article/DGXZQOGN01D120R00C26A4000000> (参照:2026-05-26).

4. 結果

1回目の課題と改善

- 重大違反の基準が甘い** →**即公開停止ルールを導入**
- 長い動画ほど不利になる** →**30秒あたりの平均ポイント制へ変更**

スコア	理由
4	0:42-0:43および1:00-1:01のシーンで、市街地での大規模な爆発と火災が描写されている
3	0:42-0:43および1:00-1:01のシーンで、爆発により建造物が炎上・破壊されている。
5	2:17-2:19のシーンで、男性が女性の顔を掴んで撮影を妨害する身体的暴力が描写されている。
5	3:21-3:22のシーンで、男性が銃器を人に向けて発砲している。
5	3:40-3:41のシーンで、男性がロケットランチャーを構えて発砲し、人や車両を標的にしていることが示唆されている。
5	3:43-3:44のシーンで、男性が別の爆発物を発射し、人や標的にしていることが示唆されている。
4	3:31-3:34のシーンで、アクション中に大規模な爆発と火災が描写されている。

↑ある動画の分析データ。爆破シーンが何度も指摘されているが、その描写になるたびにポイントが加算されており、累積得点で動画の危険性を示す本システムにおいては不利なのではないかと考えた。

2回目の課題

- 基準の細分化により、危険性と無関係な項目まで評価され**全体的に厳しすぎる**判定となった

1. 暴力・危険行為・事故に関する観点		
項目	評価項目・具体例（些細な表現から重大な描写まで）	点数
1	自殺・自傷行為の直接的・間接的な描写	+5
2	自殺・自傷行為の道具（縄や刃物）の提示・準備	+5
3	殺人や致死行為にいたるプロセスのリアルな描写	+5
4	銃器・刃物などの凶器を人に明確に向ける行為	+5
5	殴打、蹴る、踏みつけるなどの強烈な身体的暴力	+5
6	児童、高齢者、動物など社会の弱者への虐待表現	+5
▶ 重大な違反（検知時点で即時レッド判定・配信停止）		
項目	評価項目・具体例（些細な表現から重大な描写まで）	点数
1	自殺・自傷行為の直接的・間接的な描写	
3	殺人や致死行為にいたるプロセスのリアルな描写	
7	テロ行為、監禁、誘拐、人質拘束の描写	
27	実在の政治家・公人のディープフェイク偽発言	
28	命に関わる偽の医療アドバイス・健康情報の提示	

←即公開停止ルールの追加
↓点数の付け方に一貫性がなく、判定が厳しくなる基準。どこまで表現の自由を尊重するのかを考え直し、基準を修正する必要があると感じた。

84.	1秒未満で画面が切り替わる超高速編集（脳疲労の誘発）	(+3)
85.	視聴を止めさせないための無限ループ（エンドレス）構造	(+3)
86.	恐怖感、焦燥感、不安感を過度に煽る重低音BGMの継続	(+2)
87.	画面の激しい揺れや回転（映像酔い・3D酔いの原因）	(+2)
88.	奇抜すぎるネオンカラーや極端な原色の多用による視覚疲労	(+1)
89.	キャラクターの顔や目、口が不自然に巨大化する演出	(+1)
90.	倍速再生を前提とした、聞き取り不能な超高速ナレーション	(+2)
91.	画面を上下2分割し、全く異なる過激な動画を同時に流す	(+3)
92.	視聴者にカウントダウンを突きつけ、心理的猶予を奪う	(+2)
93.	過度にソワソワさせる不快なASMR（咀嚼音の誇張など）	(+1)



↑1回目
予備資料 作成した基準の一覧、分析データ



↑2回目
予備資料 作成した基準の一覧、分析データ