「ディープフェイク」の衝撃

― 最新 AI が作り出す衝撃の映像 ―

副主任研究員 柏村 祐

＜その映像は本物か＞

「百聞は一見に如かず」日本人には広く知られた有名なことわざである。何度も聞くより、一度実際に自分の目で見る方が勝る、という意味である。

しかし現実の世界に目を向けると、「一見は百聞に如かず」ということがよくある。自分が信じたいことの裏付けとなる情報を探し続け、発見すると安心し、自分が信じない情報は徹底的に無視する人をよくみかける。

このような人の習性を逆手に取った人々が使い始めたのがフェイクニュースである。

2016年の英国・EU 離脱の是非を問う国民投票、米国・大統領選の投票では SNS を通じて多くのフェイクニュースが拡散された。日本では2016年４月に発生した熊本地震直後に、動物園のライオンが脱走したというフェイクニュースが記憶に新しい。

こうした中、新たに登場したのが動画版フェイクニュースと言われる「ディープフェイク」である。ディープフェイクを一気に有名にしたのは、４月17日 YouTube にアップされた動画だ。

ホワイトハウスと思われる部屋で「トランプ大統領はまったくもって完全なばか野郎だ」と偽のオバマ前大統領は動画でスピーチしている（図表１）。

図表１　AIで作成されたバラク･オバマ前大統領

資料：YouTube(You Won’t Believe What Obama Says In This Video!)

「ディープフェイク」とは、高度な画像生成技術を駆使して合成され、偽物（フェイク）とは容易に見抜けないほど作り込まれたニセ動画の通称である。

ディープフェイクでは、オバマ前大統領のように、政治家などの著名人の顔を使って、あたかも彼らがしゃべっているかのように動画を作れてしまう。最もやっかいなのはそのクオリティである。安っぽい動画であれば人々はその動画が嘘だと気づけるが、精度が高いため、気づくのが難しい。

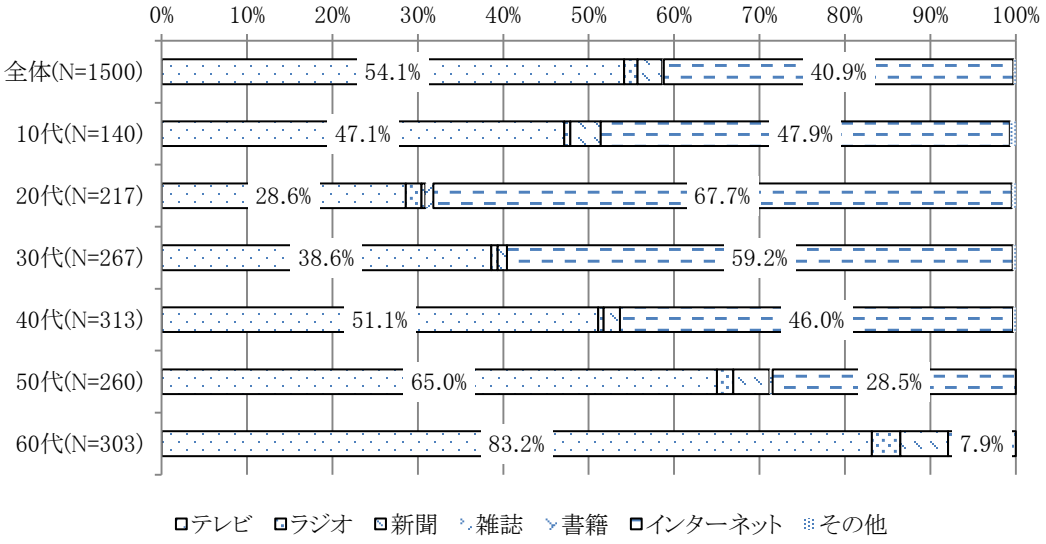
今のところ、ディープフェイクの例は、政治家におかしなことを言わせた動画などの愉快犯が作成する領域にとどまっている。

＜拡散しやすいインターネット＞

メディアとしてのインターネットの利用について、利用目的ごとに他のメディアと比較したものが、図表２、図表３である。最も利用されているメディアはテレビであり「いち早く世の中のできごとや動きを知る」（54.1%）、「世の中のできごとや動きについて信頼できる情報を得る」（57.2%）となっており、いずれの目的においても、全体でテレビが５割を超えている。

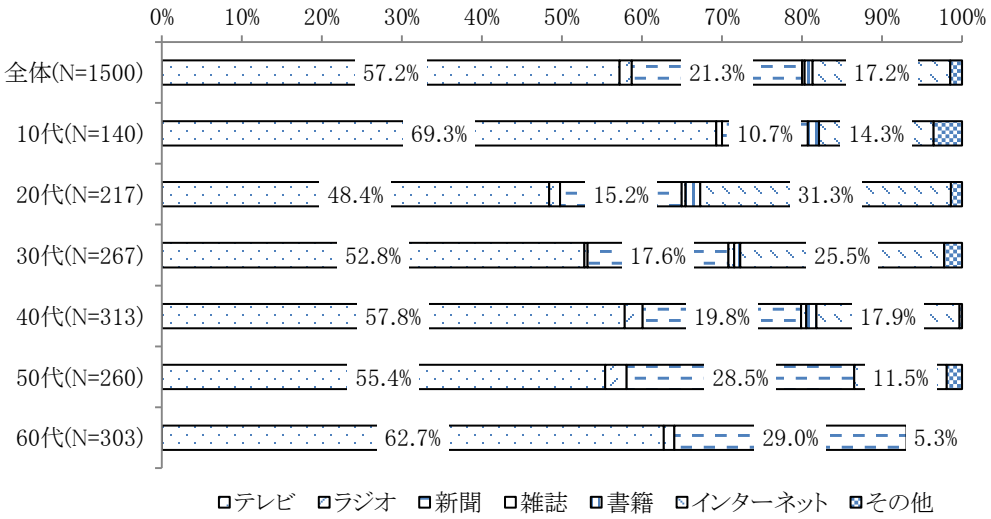
「いち早く」時事情報を得るために最も利用するメディアとしては、10代から30代までではインターネットがテレビを上回っている。前回調査と比較すると60代を除く各年代においてテレビが減少傾向、インターネットが増加傾向となっている（図表省略）。

図表2　目的別利用メディア（最も利用するメディア）
【いち早く世の中のできごとや動きを知る】



資料：総務省情報通信政策研究所「平成28年情報通信メディアの利用時間と情報行動に関する調査」より筆者作成

図表3　目的別利用メディア（最も利用するメディア）
【世の中のできごとや動きについて信頼できる情報を得る】



資料：図表２と同じ

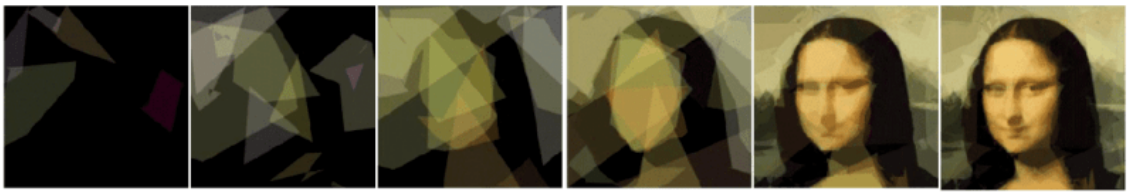
テレビや新聞等では管理する組織が情報をフィルタリング*1して世の中に発信するため、現在の日本においてはディープフェイクが発生する可能性は極めて低い。但しインターネットにおいては管理主体やルールが明確ではなく、虚偽の緊急警報を流したり、ある特定のターゲットを攻撃するための偽の動画を公開するなど、ディープフェイクを拡散しやすい状況にある。

このような状況下では例えば、虚偽の緊急警報を流したり、ある特定のターゲットを攻撃するために偽の動画を投下するなども実現できてしまうことになる。

＜敵対的生成ネットワークの活用＞

ディープフェイクは、敵対的生成ネットワーク（GAN：Generative Adversarial Network*2）を用い作成されている。敵対的生成ネットワークのしくみは生成側（動画を作るAI）と識別側（動画を判別するAI）の二つのAIによって構成されている。生成側のAIは何度でも動画を作成し、識別側のAIは動画を何度でも検証する。機械学習においては膨大な作成と膨大な検証を経て完成したものが、本物らしいディープフェイクとなる。膨大な作成と検証をするためには、作成したい対象の情報が多いほど精度が高まる。ディープフェイクの中で、大統領などの有名人が使われるのはそのためなのである。

図表4　敵対的生成ネットワーク（次第に本物に近づく様子）



資料：人工知能で生成されたデータ（create with. ai より抜粋）

ちなみに GAN の本来の用途は当然ながらディープフェイクではない。GAN は自動運転車における歩行者等の交通弱者の認識能力を高めたり、AI スピーカーのような音声認識テクノロジーの会話能力を向上させたりするのに非常に有望な AI である。

ディープフェイクを見破るのは指数学的に難しくなってきている。少ないデータセットであれば安っぽいディープフェイクしか出来ないが、GAN は常に進化しているためである。大手動画投稿サイトなどではディープフェイクの動画アップを禁止する措置を実施しているものの、判別ができないので必ずしも有効性が担保されているわけではないだろう。

一方で、ディープフェイクへの対策の研究も進んでいる。例えば海外の大学では、自らディープフェイクを作成し、真贋を識別する技術の研究が盛んだ。また DARPA（米国高等研究事業局）はすでにメディア・フォレンジック*3というプログラムで、世界中から技術者を集め、自動的に動画の真贋を判断する技術開発を進めている。

＜おわりに＞

ディープフェイクの状況はまさにコンピューターウイルスの歴史と酷似しているのではないだろうか。今広がっているディープフェイクは第一世代と言われている。第一世代のディープフェイクを見破れる AI は誕生しているがそれを上回った GAN や他の AI を駆使した第二世代のディープフェイクも近い将来誕生するだろう。ディープフェイクの誕生と真贋判定のプログラムの戦いはこれからだ。

我々は日々大量の動画を視聴している。視聴している動画が「本物」か「偽物」か自分の頭で判断し見極めることが必要なのである。

（企画総務部　かしわむら たすく）

【注釈】

*1　データをふるいにかけて分類すること。

*2　2014年にイアン・グッドフェローらによって発表された教師なし機械学習で使用される人工知能アルゴリズムの一種

*3　コンピュータやネットワークシステムのログや記録、状態を詳細に調査し、過去に起こったことを立証する証拠を集めること