嘘か誠か

「デジタルフォレンジックの父」と称されるハニー・ファリード氏が、生成AIと悪意ある改ざんコンテンツの拡大、そして企業が警戒すべきポイントを解説する。取材・執筆はブランズウィックのチェルシー・マグナントが届ける。

本記事は、英語版公開時点（2024年11月）の内容を日本語訳しております。

応用数学者でありコンピューターサイエンティストとしても研鑽を積んでいたハニー・ファリードは、1990年代後半に偶然手にした一冊の本をきっかけに、当時としてはまったくの新しい分野であった、デジタル・フォレンジクスという未知の領域へ足を踏み入れることになった。彼の旅路は、古びた図書館での出会いに始まる。

「図書館で本を借りるなんて、今ではずいぶん昔の話だと分かるでしょう」と、現在カリフォルニア大学バークレー校教授のファリードはいう。「わざわざ図書館に行って本を手に取るなんて、今思うとずいぶん風情がありますよね。」

1997年にペンシルベニア大学で計算機科学の博士号を取得した彼は、人間の知覚を対象とした脳研究のポスドクとして働いていた。図書館で貸し出しの列に並んでいたとき、何気なく近くに置かれていた『Federal Rules of Evidence（連邦証拠規則）』を手に取ったのだ。

「法律の知識は皆無だったけれど、暇つぶしだったのです」とファリードは振り返る。ぱらりと開いたページには“法廷で写真の提出を証拠として認める”と書かれていた。さらに読み進めると、裁判所がどのような写真を真正な証拠として認めるべきかが論じられており、そこには登場したばかりのデジタル写真についても触れられていた。“デジタル画像は35ミリネガと同じ真正性を有するとみなす”――その一文を目にしたとき、ファリードは「それは問題ではないか。いずれデジタルが本流となり、画像の改ざんが簡単になったとき、一体どうなってしまうのだろう」と考えたという。

2年後、初めて教職に就いたダートマス大学での授業の合間に、彼は友人の顔をテニス界のスター、アンドレ・アガシの写真に合成するという、ちょっとふざけたフォトショップ画像を作っていた。

写真：ハニー・ファリド提供

「体のサイズに合わせるために頭を少し大きくしなければならなかった」とファリードは振り返る。「そのとき、ピクセルを足したせいで画像にデジタルアーティファクトが生じたことに気づいた」という。

そして彼は、こうしたアーティファクトが画像改ざんの証拠になり得ると悟った。この二つの着想が結び付き、後に「デジタルフォレンジック」と呼ばれる分野、すなわちデジタルメディアを解析し、改ざんの有無や手口を見抜く研究へと彼を突き動かしたのである。それ以降、改ざんされた画像・音声・動画は、個人や企業だけでなく、公共の安全や民主社会の根幹を脅かす重大な問題として取り沙汰されるようになった。とりわけ、動画やデジタルツールによって人物の容姿をリアルな「操り人形」に変えてしまうディープフェイクへの対処は、いまや安全保障領域の最重要課題の一つとなっている。

しかし、ファリードによれば、2000年代初頭には関連するツールも研究も「ほとんど存在していなかった」という。「当時私たちは、論文を書き始め、このテーマについて考え始めたばかりの頃で、まだ誰も生成AIの到来を予想していませんでした。それが今では、メディアを操作・合成する技術は20年前とはガラッと様変わりしています。『先見の明があったのですね』と言われることもありますが、そもそもは、私がただフォトショップで遊んでいただけの話なのです」と語る。

現在ファリードは「デジタルフォレンジックの父」と呼ばれ、ディープフェイクや画像改ざん分野の世界的権威として知られている。映画スターや著名人はもちろん、政治家、弁護士、法執行機関、ジャーナリスト、さらにはホワイトハウスや国連までも、意図的なデジタル改変や捏造による誤情報を見抜くために彼の助言を仰いでいる。

また、彼は全米発明家アカデミーのフェローであり、初の著書『Photo　Forensics』は「この分野のバイブルになる」と『Enterprise』誌から高い評価を受けた。さらに、こうした脅威への対策を支援する企業としてGetReal Labsを設立し、本テーマにおける組織側の対策支援を行っている。

カリフォルニア大学バークレー校でファリードの教え子だったブランズウィックのチェルシー・マグナントは、ブランズウィック・レビューの取材として彼にインタビューを行い、急速に拡大する誤情報の脅威と、図書館で偶然手に取った一冊の本をきっかけに研究の道へと進んだ彼のキャリアの軌跡について語り合った。

「それがこの世界の仕組みなんだ」と彼は語る。「美しくもあり、恐ろしくもある。もしあの瞬間、少しでも早くあるいは遅くそのページを見逃していたら……私はまったく別の人生を歩んでいたかもしれない。」

何も信じることができない時代に、私たちの社会は一体どこへ向かうのか。……私たちはいま、極めて危険で不気味な世界へと、足を踏み入れつつあります。

ディープフェイクや改ざんメディアに関するファリードさんの取り組みについて、簡単にご説明いただけますか。

私たちの活動の中心となる技術面での取り組みは、画像・音声・映像を取り込み、それが改ざん・編集されたものか、あるいはAIによって完全に生成されたものかを判定する計算手法の開発になります。つまり、メディアの真正性をいかに検証するかが私たちの解く課題です。

この技術は、裁判所、報道機関、サイバー攻撃を受けるフォーチュン500企業、規制当局などで活用されています。メディアの現場では、世界各地で起きている出来事や、画像・音声・映像が改ざんやAI生成かどうかについて、記者から問い合わせが来ない日は、この1年間で一日もなかったと記憶しています。

メディア操作の進化について、近年どのような変化がみられるか、教えてください。

生成AIやディープフェイクが登場して間もない頃から、同意のない性的画像が作られるようになりました。多くは女性の顔をポルノ素材に合成し、それを恐喝や嫌がらせ、さらし、屈辱などに利用します。標的は有名人に限らず、ネットに顔写真を一枚でも載せたことのある人なら誰でも被害を受ける可能性があります。たとえばLinkedInに写真を載せていれば、その一枚だけであなたをディープフェイク動画に仕立てることができるのです。さらに、幼い子どもの写真を性的コンテンツに合成する児童性的虐待メディアまで作られており、まさに目を背けたくなる惨状です。

詐欺も巧妙化しています。小口では、息子や娘、孫、母親、父親など肉親の声を装ったディープフェイク通話がかかり、金銭をだまし取るケースが増えています。大口になると、組織ぐるみの手口で偽装した相手に数千万ドル規模の資金を送金させられるなど、被害は甚大です。

採用面接では、ライブのビデオ通話で候補者と話しているつもりでも、映っているのは別人で、結果的に社内にマルウェアを仕込むハッカーを招き入れてしまう、そんな事例がすでに何度も起きています。

生成AIの登場に伴い、偽情報や選挙干渉も確実に増えています。こうした問題がなくなる気配はなく、むしろ悪化の一途をたどるでしょう。

さらに深刻なのは、読んだこと、見たこと、聞いたことがすべて偽物かもしれない世界では、何も本物である必要がなくなる点です。昨年、バイデン大統領が次期大統領選から撤退した際にハリス副大統領が記者会見を開きましたが、COVIDに感染していたバイデン氏は電話で4分ほど参加し、副大統領を紹介しました。とこ

ろが一部の人々は「あれは偽物で、彼はもう亡くなっている」と主張し、議員まで巻き込んだ調査要求の陰謀論が広がりました。なぜそんなことが起きるのか。必ずしもそれが現実である必要がないからです。

何も信じることができない時代に、私たちの社会は一体どこへ向かうのか。メディア、政府、そして科学者への信頼がすでに揺らいでいるところに、さらに追い打ちがかかっています。私たちはいま、極めて危険で不気味な世界へと、足を踏み入れつつあります。悪意ある者は偽情報を作り出すだけでなく、それをネット上に絨毯爆撃のように無差別にばらまき、その情報はソーシャルメディアのアルゴリズムによって増幅されてしまいます。

この環境で人々を守るために、企業が取るべきベストプラクティスについて少しお聞かせください。

もし貴社がフォーチュン1000企業であれば、こんな事態を危惧する必要があります。あなたの会社のCEOが、街中で子犬を蹴っている動画を誰かが捏造し、X（旧Twitter）に投稿して数百万回再生されるかもしれません。そんな映像が出回れば、企業イメージは大きく傷つき、対処は極めて困難です。人は一度見た動画を忘れられません。あるいは、CEOが「利益が5%減少した」と語る偽の決算説明会の配信動画を作られ、あなたが状況を把握するよりも前に株式市場で数十億ドルが動いてしまうこともありうるのです。社内の誰かが CEO や CTO、CFO からだと思い込んだ電話で「これをやってくれ、これを教えてくれ、この情報を渡してくれ」と指示を受けるケースも起こり得ます。音声クローンを悪用したパスワードリセット詐欺も既に報告されています。ソーシャルエンジニアリング、すなわち人間をだましてセキュリティ上のミスを誘発する手口は現実の脅威であり、ディープフェイクはこれらの攻撃を社内外でさらに強力にしてしまうのです。

では、どう対処すればよいのでしょうか。この分野でできるのは排除ではなく緩和策です。誰かがあなたを害そうと決めれば、実際に害を与えることができてしまいます。ただし、その被害を抑え、ダメージを最小限に抑えることは可能です。

もし私が組織や政府のトップであれば、公開するあらゆるコンテンツに自分のデジタル署名を付けます。決算説明会の音声や画像、ビデオインタビューなど、すべてです。そうしておけば、決算説明会だとか写真だとか演説だとかを名乗るものが出回っても、署名がなければ「偽物だ」と即座に判断できます。これが第一の対策です。

第二の対策は、テーブルトップ演習を実施することです。どのように対応するのか、最初に誰に連絡し、次に誰に連絡するのか、X（旧Twitter）やFacebookから偽情報をいかに迅速に削除するのか、犯人をどう突き止め責任を取らせるのか。こうした一連の手順を事前にシミュレーションしておくのです。なぜなら、攻撃者が逃げ切れば、同じことが必ず繰り返されるからです。組織内で誰が担当するのか、CISO（最高情報セキュリティ責任者）なのか、広報なのか、多くの組織では、この責任の所在すら曖昧なのが現状です。

プレスリリースは即座に発信しなければなりません。こうした事態への対応には、数分しか猶予がありません。ソーシャルメディアの投稿の「半減期」は、ものの60秒程度で訪れます。言い換えると、全インタラクションのうち、おおよそ半分は最初の1分で生じるのです。対応に数時間や数日をかける余裕はなく、対処の猶予があるのはせいぜい数分なのです。

では、この機能は組織内のどこに置くのがベストでしょうか。

良い質問です。現状では　CISO（最高情報セキュリティ責任者）がこの領域を担うケースが増えており、私もそれが妥当だと考えています。ただし彼らの専門分野とは言い難く、詳しい人は多くありません。そのため私は　CISO　はもちろん、CEO とも時間をかけて議論しています。

私たちはこれまで、操作されたメディアを「不気味の谷」という視点から語り合ってきました。その点について説明していただけますか。

「不気味の谷」という言葉は、もともと人と直接触れ合うヒューマノイドロボットをつくるロボット工学の分野から生まれましたが、いまでは画像・音声・映像にも使われています。アニメのようにデフォルメされた映像は楽しく観られる一方、人間に近づきながらも完全には人間らしくない段階に差しかかると、私たちは強い違和感を覚え、「不気味だ」と感じます。

ところが現在、人を写した偽造画像はその「不気味の谷」をすでに越えています。きわめてフォトリアリスティックで、見ても不自然さや不快感がありません。画像を見ただけでは、本物の人物かどうかを確実に見分けるのはほぼ不可能です。音声、つまり話し声もほぼ谷を抜けつつあり、偽物か本物かを判定する正答率は、もし当てずっぽう（ランダム）に回答すれば50%であるところ、現在は約65%と、まだランダム水準よりわずかに高い程度にとどまっています。

動画の状況はまだ玉石混交といえるでしょう。1年前に話題になった「ウィル・スミスがスパゲッティを食べる」映像を思い出してください。

きっとこんな事態が起こるでしょう―誰かがあなたの会社のCEOが、街中で子犬を蹴っている動画を捏造し、X（旧Twitter）に投稿して数百万回再生されるかもしれません…あるいは、CEOが「利益が5%減少した」と語る偽の決算説明会を作られ、あなたが状況を把握するよりも前に株式市場で数十億ドルが動いてしまうこともありうるのです。

当時は笑えるほど奇妙でしたが、今もかなり改善されたとはいえ、まだ完璧ではありません。一方、人物の顔を別の映像に重ねるディープフェイク動画はすでに高品質ながら、やはりこちらも完全ではないのです。ただ半年、1年、1年半もすれば、こうした技術はますます高度になり、低コストで手軽になり、利用も増えていくでしょう。

生成AIは、わずか数年という短期間で「ひどい出来」から「本物と区別がつかない」レベルへ進化しました。動画も同じ道をたどると私は考えており、その実現まであと1年から1年半ほどだと思います。

では、先ほど挙げた諸問題にとって、この技術進歩は何を意味するのでしょうか。

状況はさらに悪化していきます。望みはただ一つ、対策がそのスピードに追いつくことです。スパムもウイルスも深刻化しますが、対策技術も進化する。すべてが同時に加速していくことが唯一の希望です。

ところが、ここシリコンバレーでは、ベンチャーキャピタルが生成AIに投じる資金と、防御や介入に投じる資金とでは桁が違います。生成AIには数十億ドルが流れ込む一方、防御系、つまり私たちのような企業にはせいぜい数百万ドルです。

防御は難しく、収益性も低い。そのため、VCの資金力も人材も学術的リソースも、防御側はやや劣勢に立たされています。これは憂慮すべき事態です。いつかバランスが取れることを願っていますが、その兆しは見えません。むしろ、状況はさらに厳しくなるでしょう。

それでも希望はあります。社会全体の意識向上、私たちが今こうして交わしているような対話、一定の規制圧力、そして優れた技術。これらが相乗効果を発揮することで、リスクの一部は必ずや軽減できるはずです。

規制面の現状は、いまどこまで進んでいるのでしょうか。

昨年10月、バイデン大統領は生成AIから予測AIに至るまで、AI全般を対象とした大統領令を発令しました。これを受け、米国国立標準技術研究所（NIST）内に「AIセーフティー・インスティテュート」が設置され、これらの課題への対応を担っています。

私は連邦議会関係者と頻繁に話をしていますが、FTC、FCC、司法省、NSA、CIA、FBI、さらには行政府も立法府も、AI問題を考えていない部門は一つもありません。現状はやや断片的で非効率なため統合が求められますが、AI、特に生成AIがもたらす新たな世界について、皆が真剣に向き合い始めています。

ソーシャルメディアの投稿の「半減期」は、ものの60秒程度で訪れます。…対応に数時間や数日をかける余裕はなく、対処の猶予があるのはせいぜい数分なのです

ホワイトハウスは国際連携にも力を入れており、オーストラリア、カナダ、英国、EUなどの同盟国と協力して包括的に議論を進めています。実際、課題の約95%は米国外にあります。それでもテクノロジー業界は米国中心に偏りがちで、たとえばコンテンツモデレーションが最も必要とされる地域ほど、担当チームに現地語を話せる人が一人もいない。そんな状況が存在するのです。

「どの国が最もうまく取り組んでいるか」と聞かれれば、私は迷わずオーストラリアだと答えます。AI規制、ソーシャルメディア規制、独禁法のような規制など、あらゆる面で群を抜いています。eSafetyコミッショナーのジュリー・インマン・グラントはまさに圧倒的なリーダーで、私は皆に「オーストラリアの取り組みを見てほしい」と勧めています。EUはやや規制強化に走り過ぎている面もありますが、その方向性は評価できます。英国にもオンライン安全法がありますし、バイデン大統領の大統領令（法的拘束力はありませんが）も総じて良い内容だと思います。

私が現在、先導役として注目しているのは以上の4つです。加えて、カリフォルニア州もかなり健闘しています。州レベルで規制を行うなら、テック企業とVC資金の大半が集まるカリフォルニアこそが担うべきでしょう。同州の法案で示される文言には好感が持てますが、VCコミュニティからは激しい抵抗を受けています。

そのお話を伺うと少し怖くなります。正直に申し上げると、私はテクノロジーに対して少々楽観的なところがありまして…

そうですか、それで言いますと、私はどちらかというと反対の立場にいます。「インターネットは面白い実験だった、もう終わりにしよう」と思ってしまうような日もあります。でも、少しでも希望がなければこの仕事はしていません。私たちの取り組みは必要ですが、それだけでは十分ではありません。必要なのは、私たちの下流にいるテック企業を含む人々、そして上流にいる規制当局が関心を持つことです。

ありがとうございました。素晴らしい機会でした。

いつもながら、お会いできて嬉しかったです。ありがとうございました。

チェルシー・マグナント： チェルシーはブランズウィック・グループのAIクライアント・インパクト・ユニットを率いる。キャリアをCIAでスタートし、米国上級政策立案者の地政学課題対応を支援。前職では、Google にてテック政策戦略に携わった。