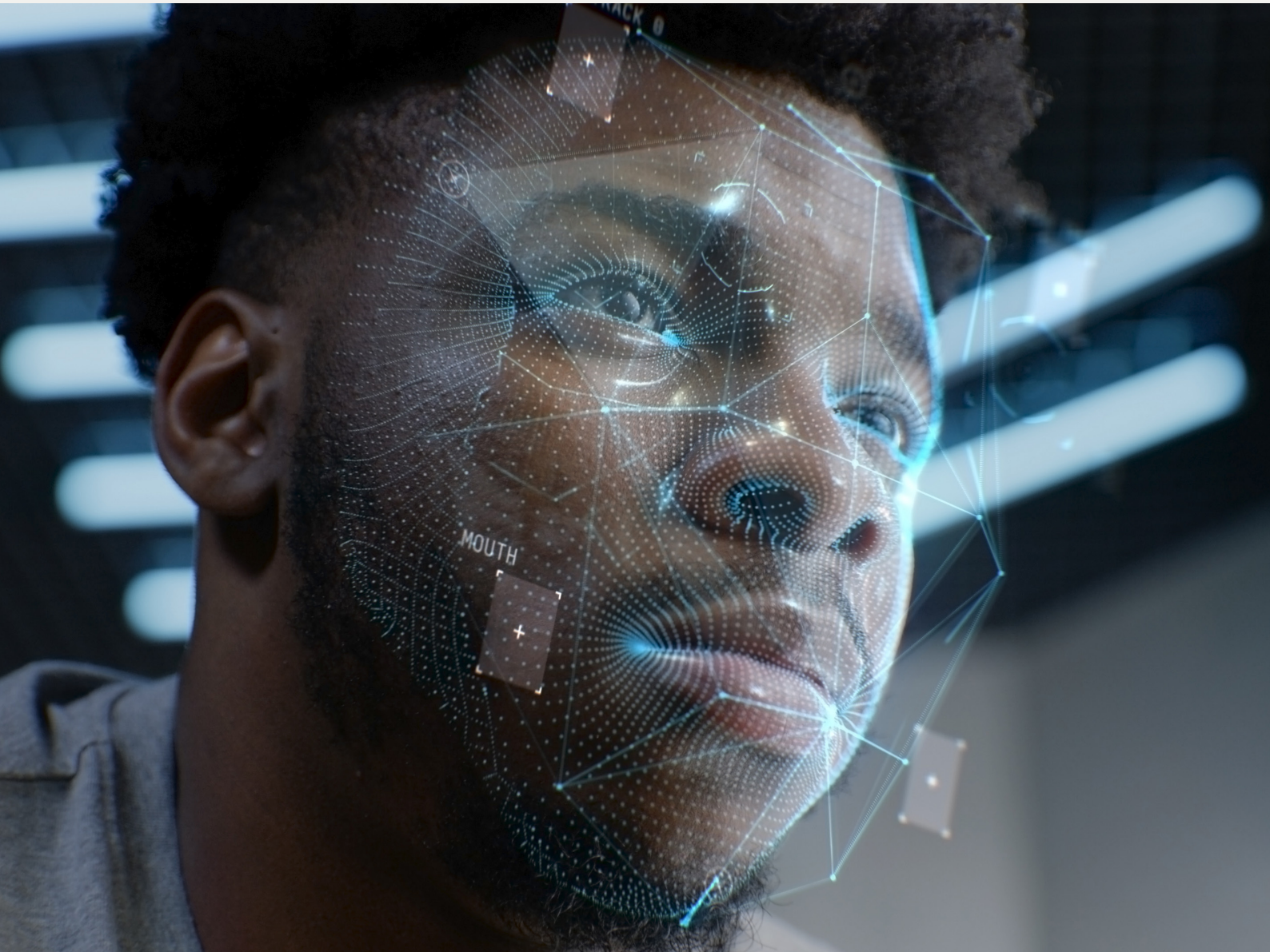
kyndryl

ディープフェイクから自社を守る方法

クリス・ラブジョイ著
キンドリル セキュリティ&レジリエンシー担当グローバルリーダー



重要なポイント

AIを活用した詐欺は、今や経営レベルのリスクとなっています。ディープフェイク攻撃を認識して防御する能力は、今後企業にとって決定的なスキルとなります。

人間の「信頼（人間の心理的な弱点）」が新たな攻撃対象となっています。サイバー犯罪者はソーシャルエンジニアリングを産業化しており、能動的な検証と多層防御が不可欠となっています。

非公開の共有知識が究極の防御策です。ディープフェイクが公開データから構築されている場合、内部だけで共有している情報に基づいて検証することが、攻撃を阻止するのに役立ちます。

はじめに

AIを活用した詐欺は、あらゆる企業にとって顕在化し増大する脅威となっています。

脅威は根本的に変化し、生成AIはソーシャルエンジニアリングを手作業の技から産業化されたプロセスへと一変させました。サイバー犯罪者は今や、フィッシング、ビッシング（ボイスフィッシング）、ディープフェイク（音声・映像・文章のなりすまし）攻撃を、これまでは想像できなかった規模と巧妙さで実行できるようになりました。

リスクは高く、最近のデータによると、ディープフェイク攻撃は数十万ドル規模の損失をもたらし、数百万ドル規模の詐欺に発展する場合もあります。また、経済的な損失だけでなく、ステークホルダーの信頼、顧客の信頼や信用を損なうといった企業の評判を悪化させる可能性があります。さらに、ディープフェイク攻撃は、驚異的なスピードで増加しています。2025年のガートナーの調査によると、62%の組織が過去12カ月間に少なくとも1回のディープフェイク攻撃を経験しています。

経営陣が誤った方向に導かれるのは時間の問題と言える状況において、被害の大きさを左右するのは防御の堅牢性です。サイバーセキュリティ、先進技術、人的ファイアウォール、強固なガバナンスを包含する多層的フレームワークは、もはや選択肢ではなく、必須要件です。

詐欺の産業化

従来のフィッシング攻撃は、文法の誤りや画一的な挨拶文で正体が判明したり、スパムフィルターにひっかかることが多くありました。しかし、今日の悪意ある攻撃者は、AIツールを用いて、声、顔、文章のスタイルを完璧に模倣することができます。

AIは、攻撃開始前の偵察からコンテンツ作成までの攻撃ライフサイクル全体を自動化します。ソーシャルメディアや企業のウェブサイトから公開データを収集して詳細なターゲットプロファイルを構築し、そのデータを活用することで、特定のプロジェクト名、同僚、個人的な出来事までを盛り込んだ、極めて個別化されたメッセージを作成します。

サイバー犯罪者は、高精度の音声クローン技術を用いて経営幹部になりすまし、財務担当者に電話をかけ、緊急かつ正当と思われる依頼を行い、送金を指示するというような音声なりすまし詐欺（ビッシング）を行うことができます。わずか数秒の音声から、驚くべき精度で音声を複製することができるのです。

デジタルなりすましの最前線から得られる教訓

実際に発生したインシデントを分析することで、この緊迫した局面を乗り切るための実践的な教訓を得ることができます。

2024年に発生した現代的なソーシャルエンジニアリングのインシデントでは、金融企業の従業員が偽のディープフェイクを用いたビデオ会議によって騙され、2500万ドルの送金を行いました。この詐欺が成功した要因は、その多層的な複雑さにありました。攻撃者は1人だけでなく、CFOや他の上級管理職のディープフェイクでビデオ通話全体を演出したのです。ここで覚えておくべき重要なポイントは何でしょうか。「同僚に確認する」といった標準的なアドバイスは、デジタル上の同僚すらも偽物である場合、もはや通用しません。

この事例やその他の多くの事例において、企業はデジタル通信を誤って信頼し、高リスク取引に対するバックアップの確認を行っていません。デジタルメッセージが偽物である可能性を前提とした、完全なプロセスが存在しないのです。

しかし、事例研究から、AIを活用した攻撃の根本的な弱点も明らかになっています。それは、公開されているデータから構築されているという点です。ディープフェイクは、公開された人物のみを複製することができます。

このため、最も強固な防御策は、攻撃者が収集または合成できない非公開の共有知識に基づく検証手順を構築することです。これは、高級車メーカーであるフェラーリに対する最近の攻撃によって実証されています。攻撃者の計画は、ディープフェイクが持ち得ない知識に基づいて従業員が判断したことにより阻止されました。

レジリエントな防御のための枠組み

今や、包括的な防御戦略は、技術的な要件以上に事業継続と財務リスク管理にとって極めて重要となっています。

第 1 の柱：
先進なエンタープライズアイデンティティ認証

高度な脅威は基本的な脆弱性を悪用して成功することが多く、サイバーセキュリティの予防策が最初の防御となります。これには、攻撃者がなりすましを画策するために必要な最初の足がかりを防ぐパッチ管理や多要素認証（MFA）が含まれており、リスクを99%以上低減することが可能です。侵害が発生した場合は、ゼロトラストの原則とアイデンティティアーキテクチャにより、攻撃者が損害を与える能力をさらに制限することができます。

第 2 の柱：
AI で AI と戦うための高度な対策

第2の柱は、AIでAIと戦うための先進テクノロジーの活用です。AIを活用した検知ツールは、生成されたコンテンツにおいて、人間には感知できない微妙な統計上の異常を特定することができます。リアルタイムの視覚および音声分析、さらには生理学的分析を通じて、これらのツールは映像フィードにおける血流の欠如や自然な人間の微妙な表情の欠如を検知し、ディープフェイク技術であることを見抜くことができます。

第 3 の柱：
心理的なセキュリティエンジニアリング

技術だけでは不十分であるため、第3の柱として、組織は人的ファイアウォールを強化する必要があります。リーダーは、CEOからのものを含むすべての高額取引や機密性の高い依頼について、事前に合意された動的な「秘密の質問」またはパスフレーズによる検証を義務付けなければなりません。公開され、誰もがアクセスできるデジタルシステムには一切記録されないことが必要です。また、AIを活用したシミュレーションプラットフォームにより、複製された経営陣の声によるなりすまし電話など、現実的なトレーニング演習を実施することで、従業員に立ち止まって確認する習慣を身につけさせることができます。

第 4 の柱：
ディープフェイク対応マニュアル

最後に、組織はガバナンスに目を向け、ディープフェイクに特化したインシデント対応マニュアルを策定する必要があります。迅速で組織的かつ多面的な対応を確保するため、このマニュアルには、フォレンジック検証、技術的な封じ込め、危機管理広報、法執行機関との連携に関する、事前定義されたワークフローを含める必要があります。

チーム/役割	緊急対応（最初の60分間）	フォローアップ対応（1〜24時間以内）
セキュリティ運用	システムを隔離 悪意のあるIPアドレスをブロック フォレンジック分析を開始	詳細なログ分析を実施 関連するIOCを追跡
IT/インフラストラクチャー	侵害された認証情報をリセット システムの完全性を確認	関連する脆弱性の修正を加速
法務/コンプライアンス	弁護士に相談 証拠を保全	規制当局への通知を評価コンテンツの削除を要請
企業広報/PR	危機対応計画を発動 一時的な声明を発表	詳細な公式声明を作成・発表 メディア対応を行う
ファイナンス部門	銀行に連絡し、不正な送金を停止または取り消す	完全な監査を実施 送金プロトコルを強化

AI 軍拡競争に勝つ

生成AIの出現は、サイバー犯罪者とグローバル企業の終わりのないテクノロジー競争の始まりを告げました。AIを駆使した詐欺の脅威に対抗するためには、リーダーは受動的なセキュリティ態勢を打開し、能動的に真正性を検証しなければなりません。企業のセキュリティの未来は、あらゆる攻撃を阻止する能力ではなく、現実そのものが主な攻撃対象となった世界において業務の完全性を維持する能力によって定義されるでしょう。

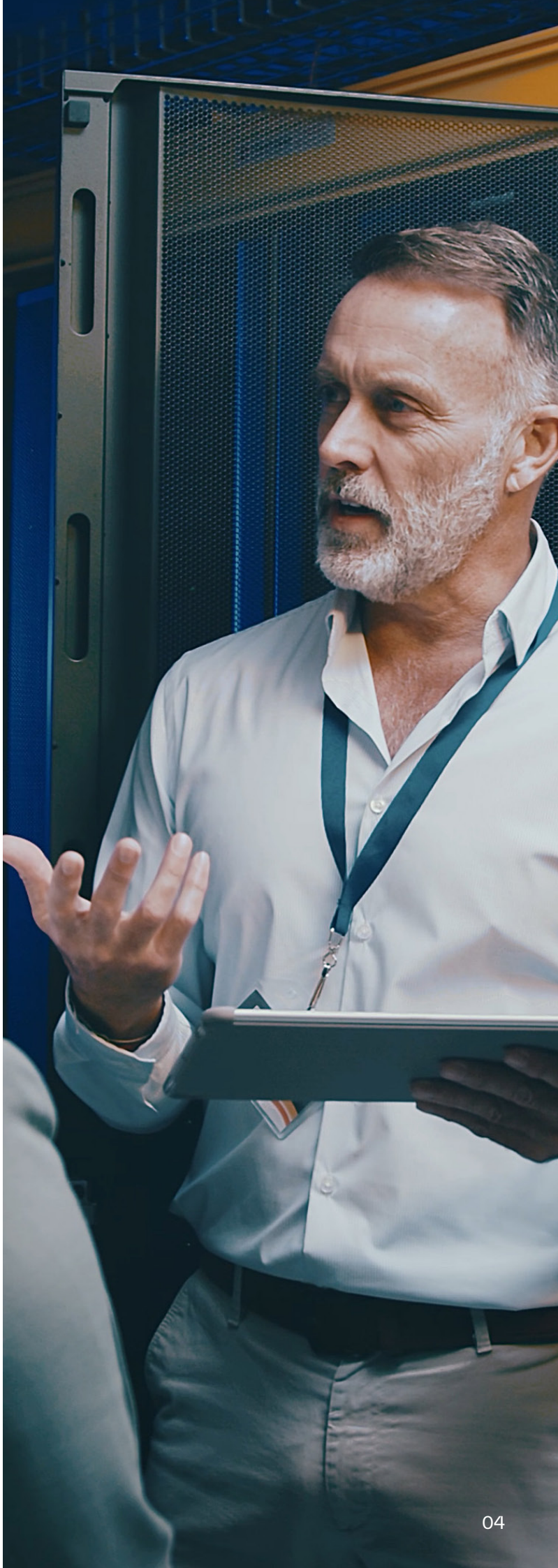
この新たな脅威には、優れたツール以上のものが求められ、組織が通信の正しさを確認し、リスクを管理し、対応する従業員を訓練する方法を根本的に見直すことが必要です。AIを駆使した詐欺の時代において、企業のレジリエンスを定義する4つの戦略的優先事項は次のとおりです。

ゼロトラストアーキテクチャーを採用する：境界ベースのモデルから、ユーザーや通信が本質的には信頼できないことを想定したモデルへ移行します。すべてのアクセス要求と機密性の高いトランザクションを継続的かつ厳格に検証します。

多層防御に投資する： 万能な解決方法は存在しません。堅牢な防御システムは、ネットワークセキュリティ、エンドポイント保護、コンテンツ検証、強固な認証プロトコルを統合し、1つの層が突破されても別の層によって防御できるようにします。

健全な懐疑の文化を育む：最もレジリエントな組織とは、従業員一人ひとりが、一見信頼できるものであっても通常とは異なる依頼や緊急の要請に対して立ち止まり、検証する習慣がついている組織です。

公開データのフットプリントを制限する：ディープフェイクは、公開データを基に作成されます。経営陣や主要な従業員に対し、オンラインで共有する高品質な音声・動画コンテンツには留意するよう助言してください。トレーニングデータを制限することで、攻撃者は精度の高いフェイクを作成することが困難になります。



kyndryl