Діпфейки

Вже нічому не можна довіряти, але виявлення підробленого контенту теж потроху автоматизують.

Діпфейки змінюють кіберзлочинність. Підробки голосу й облич за допомогою штучного інтелекту досягли безпрецедентної достовірності, що не лише змінює соціальну інженерію, а й становить загрозу для автоматичних інструментів на кшталт систем управління автентифікацією та доступом. Понад те, вже на підході автономні агенти, які самостійно втиратимуться в довіру від імені замовника, досконало імітуючи людину.

Засилля діпфейків вимагає зміни стратегій захисту і швидкого виявлення згенерованого ШІ контенту. Зокрема боротися з ШІ має також ШІ. Цікавимося, які

можливості відкриває ШІ-генерація голосу та відео перед злочинцями, як виглядають атаки і що можна зробити, щоб їм протидіяти.

Темний бік ШІ

Як пояснює Мохаммед Халіл, архітектор кібербезпеки з компанії DeepStrike, що спеціалізується на пентестах, діпфейк являє собою будь-який синтезований контент (зображення, відео чи аудіокліп), у якому образ або голос людини видозмінені або заміщені за допомогою штучного інтелекту — по суті, цифрову ляльку, яку, втім, не відрізнити від реальної особи. З технічної точки зору це історично досягається за допомогою механізму машинного навчання — т. зв. генеративної змагальної мережі (GAN), яка містить два ШІ: «генератор», який створює фейки, і «дискримінатор», який намагається їх викрити (**рис. 1**). Так обидва безперервно навчають один одного, поки фейк не стає настільки переконливим, що його неможливо виявити. Сучасніші моделі (трансформери і дифузори) створюють ще якісніші фейки.

Є цікава інфографіка (2022 рік) від американської агенції передових оборонних проєктів DARPA, що показує, які технології можуть використовувати гравці різного рівня: від індивідуальних зловмисників з генеративним ШІ до держорганів зі зйомками рівня Голівуду (**рис. 2**).

Identity Management Institute — організація, яка проводить навчання

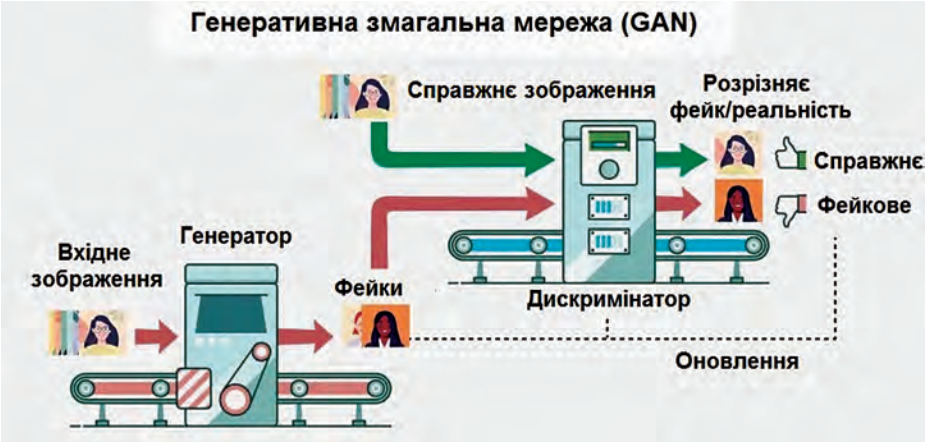

Рис. 1. Генеративна змагальна мережа (джерело: портал GeeksforGeeks.org)

і сертифікацію у сферах управління обліковими записами, управління ризиками і дотриманням законодавства (комплаєнсу) — зазначає у своєму блозі, що діпфейки становлять значний ризик для управління обліковими даними і доступом, оскільки хакери можуть удавати легітимних користувачів і викрадати секретну інформацію. Діпфейки руйнують довіру до процесів автентифікації за допомогою розпізнавання облич і голосу, розмиваючи впевненість у відповідних системах і підвищуючи ризики несанкціонованого доступу.

У традиційних фішингових схемах хакери видурюють персональні дані через електронну пошту. Діпфейки вносять новий рівень реалізму: злочинці можуть достатньо переконливо удавати колегу або начальника, щоб змусити жертву розкрити чутливу інформацію або санкціонувати платіж чи переказ.

Окрім фінансових збитків, це загрожує репутаційними втратами. Фейкове відео, у якому хтось із керівників компанії, наприклад, робить образливі заяви або поводиться неетично, може серйозно вдарити по бренду і призвести до падіння акцій або й судових позовів.

Особливо вразлива до таких атак біометрична ідентифікація — ключова технологія, що використовується в сучасних системах управління обліковими даними і доступом (IAM). Імітуючи голос, обличчя і рухи користувача, зловмисники можуть отримати доступ до захищених ресурсів. Достовірність і якість діпфейків вже досягла рівня, коли ці системи не здатні відрізнити справжню людину від підробки.

Для деяких секторів, як-от державного, фінансового і медичного, атаки з використанням діпфейків особливо небезпечні в силу їхніх наслідків. Наприклад, фейкове розпорядження від посадовця може призвести до коригування політики або до нецільового виділення коштів. У сфері охорони здоров'я діпфейки можуть вилитися в зміну даних про пацієнта або імперсонації лікаря.

Діпфейки є не лише загрозою для компаній, а й потужним інструментом дестабілізації суспільства. З їхньою допомогою можна формувати громадську думку, провокувати невдоволення і навіть насильство. Діпфейкове відео, на якому політик поширює неправдиву інформацію під час кризи, здатне спровокувати паніку або ринкову нестабільність.

Рис. 2. Творці діпфейків і доступні їм ресурси (джерело: DARPA via OWASP)

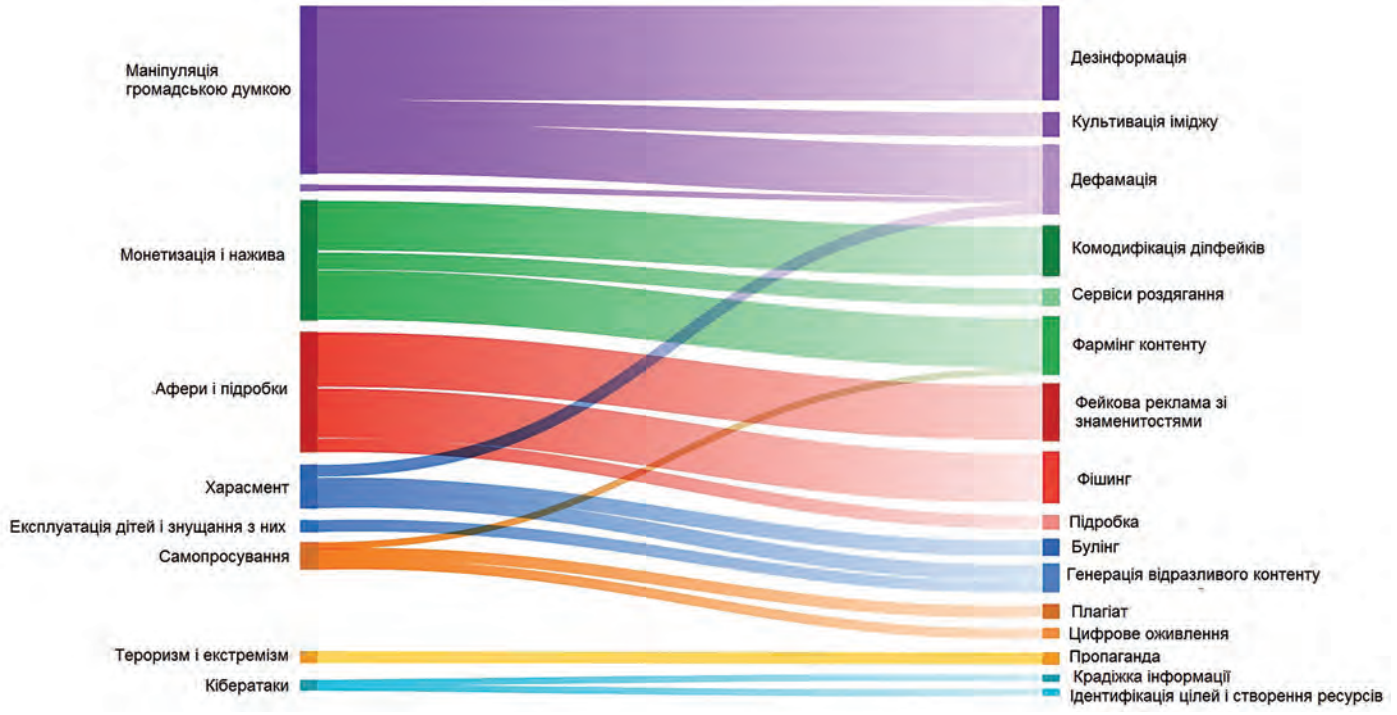

Рис. 3. Використання генеративного ШІ в зловмисних цілях (джерело: Google DeepMind, 2024)

Імітовані промови або дії політичного характеру можуть використовуватися ворогами для підриву довіри до демократичних інституцій і впливу на вибори. Наприклад, у 2024 році під час праймеріз Демократичної партії у штаті Нью-Гемпшир виборці отримували голосові дзвінки з записом, на якому нібито Джо Байден закликав не голосувати. Винних знайшли, але реалістичність підробки голосу і тону президента збентежила виборців і викликала тривогу щодо захищеності комунікацій під час виборів.

Окрім усього іншого, діпфейки справляють глибокий психологічний ефект. Наприклад, жертви цифрової порнографії зазнають потужної емоційної травми, яку зазвичай ще більше посилює вірусний характер поширення відео. У 2024 році в одній зі шкіл австралійського штату Вікторія учні створили близько 50 фейкових відвертих фото своїх однокласниць і вчителів, що викликало обурення у спільноті і призвело до розслідування та арештів. Як стверджує Халіл (хоча, мабуть, до цих оцінок треба ставитись з обережністю), такий контент, відомий як «інтимні фото, зроблені без згоди», а також як «діпфейкова порнографія» і «порно помсти», складає 96–98% усього обсягу діпфейкових відео в Інтернеті, що вже призвело до появи низки законодавчих ініціатив з метою обмеження і заборони цього неподобства.

Населення ж дедалі більш скептично налаштоване щодо правдивості контенту, який йому показують. Ця втрата довіри, яка ще носить назву «дивіденд брехуна», служить інтересам тих, кому вигідно реальні події виставляти як фейкові, що ще більше дестабілізує суспільний дискурс.

Вибухове зростання

У 2025 році компанія IBM вперше включила дані про атаки з використанням ШІ у свій звіт Cost of a Data Breach. Серед іншого дослідники зазначають, що 16% всіх організацій, які було зламано, зазнали атак з використанням ШІ. У більшості цих атак було задіяно соціальну інженерію: фішинг (37%) і діпфейки (35%).

Google має власну статистику: у 2024 році компанія опублікувала звіт з назвою Mapping the Misuse of Generative AI, складений на основі близько 200 повідомлень у ЗМІ. За період дослідження більшість випадків зловмисного використання ШІ були спробами вплинути на громадську думку, вчинити якусь аферу або заробити (**рис. 3**). При цьому дослідники зауважують, що є й деякі приклади зловживання, які не є завідомо зловмисними, але все ж ставлять етичні питання: наприклад, коли урядовці раптом починають говорити мовами своїх виборців, не розкриваючи, що користуються ШІ, або коли активісти використовують згенеровані голоси жертв стрілянини для просування реформи закону про володіння зброєю.

Ще більше даних наводить у своєму дописі вже згаданий Мохаммед Халіл. За його підрахунками, ще у 2024 році атаки з використанням дипфейків ставались кожні 5 хвилин. Число атак подвоюється що два місяці, тобто не просто зростає, а з прискоренням (експоненційно). «Сам по собі обсяг синтетичного контенту в поєднанні з перетворенням його на зброю задля фінансової вигоди створює ідеальний шторм як для людей, так і для бізнесів», — зазначає дослідник.

Халіл пише, що зростання обсягу фейкового контенту з 500 тис. у 2023 році до 8 млн у 2025-му (**рис. 4**) і вибухове збільшення кількості фейкових відео являє собою «вірусне розмноження» і залишає далеко позаду темпи поширення інших кіберзагроз. Лише у 2023 році було зафіксовано удесятеро більше інцидентів з залученням діпфейків, ніж попереднього року, що свідчить про широке освоєння цієї технології зловмисниками.

Хоча діпфейки використовуються

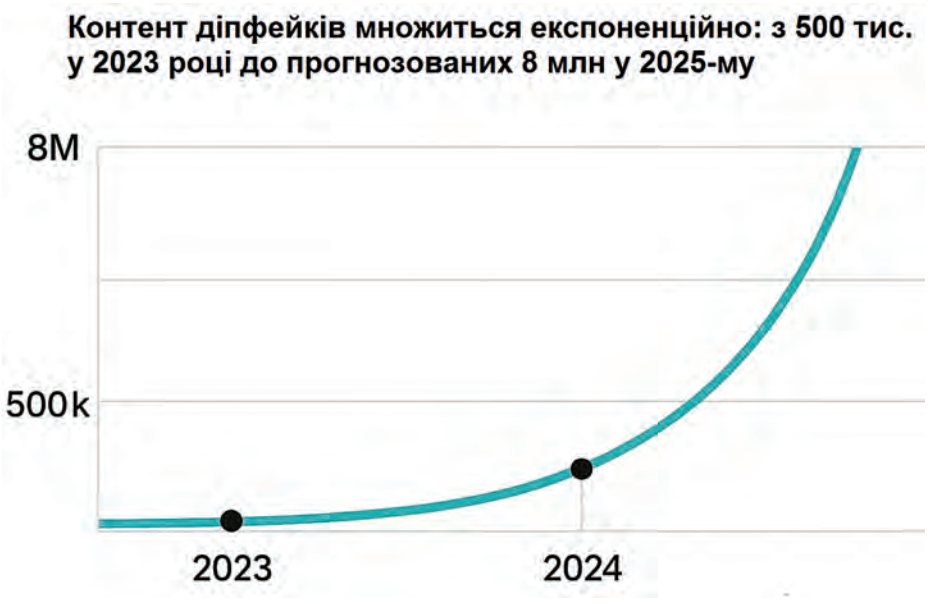

Рис. 4. Експоненційне зростання числа діпфейків (джерело: Мохаммед Халіл, DeepStrike)

в різних цілях, пише Халіл, їхнім головним зловмисним застосуванням є шахрайство. У 2023 році кількість спроб імперсонації зросла на 3000%. Фінансові збитки від цих атак теж вражають. У 2024 році середній інцидент за участі діпфейків «коштував» $500 тис. Великі підприємства втрачали до $680 тис. Deloitte прогнозує, що в США загальні збитки від шахрайства з залученням ШІ виростуть від $12,3 млрд у 2023 році до $40 млрд у 2028-му: щорічне збільшення на 32%. З точки зору географії загрози розподілені нерівномірно. Головною ціллю є Північна Америка, де з 2022-го по 2023-й рр. обсяги фроду зросли на 1740%, а фінансові збитки лише у 1 кварталі 2025 року склали $200 млн.

Найбільш поширеними, через свою демократичність, є атаки з використанням клонованого голосу. За даними McAfee, у 2024 році з голосовими аферами стикався кожен четвертий дорослий, а кожного десятого намагались «розвести» персоналізованою атакою. Халіл зазначає, що «бар'єр доступу до таких атак цілковито обвалився». Аферистам достатньо трисекундного зразка голосу, щоб клонувати його з достовірністю у 85%, а зразки можна знайти на YouTube, в подкастах, соцмережах, корпоративних вебінарах тощо. При цьому технологія неймовірно дешева і доступна: згаданий вище нью-гемпширський дзвінок від Джо Байдена був створений за 20 хвилин і обійшовся у 1 долар.

Інші приклади

У березні 2022 року з'явилося відео, на якому нібито Володимир Зеленський наказував українським військам скласти зброю. Фейк був поганої якості, і користувачі його швидко викрили, але президент мусив окремо виступити зі спростуванням.

У січні 2024 року на різних платформах, зокрема X і Reddit, з'явились фейкові порнографічні фото співачки Тейлор Свіфт. X не поспішала їх прибрати, і одне фото набрало 47 млн переглядів, перш ніж його видалили. Хештег, пов'язаний з цим вкидом, увійшов у тренди X, що ще більше сприяло поширенню. The Guardian писав, що лише після того, як фанати Свіфт мобілізувались і почали масово репортити зображення, платформа прибрала їх і заблокувала пошук. Багатьом іншим жертвам, навіть знаменитостям, пощастило значно менше. При цьому компанії на кшталт Discord вказують, що в них є модератори, а X — що має «нульову терпимість» до публікації оголених фото без згоди, але якщо зображення побачили мільйони людей перш ніж його видалили, це звучить дещо порожньо, зазначає видання.

У липні 2024 року на YouTube з'явився стрім нібито Ілона Маска, який на практиці являв собою закільцьований згенерований ШІ аудіозапис тривалістю у майже 7 хвилин. На ньому мільярдер говорив про американську політику, майбутні вибори і майбутнє криптовалюти, а ще закликав просканувати QR-код і перейти за «безпечним посиланням», щоб «легко і просто» подвоїти свої криптовалютні статки. Стрім проіснував 17 годин і в певний момент мав 104 тис. глядачів, йдеться у блозі компанії Pindrop. Акаунт, на якому розмістили відео, був активний з 2022 року, мав значок підтвердження, налічував 163 тис. підписників і понад 34 млн переглядів, на перший погляд він нагадував офіційний канал Tesla.

Можливо, найбільш яскравий приклад діпфейкової атаки трапився у лютому 2024 року, коли працівника фінансового відділу глобальної проєктувальної фірми Arup обманом змусили перерахувати $25,6 млн на рахунок, що належав зловмисникам. Працівник фінансового відділу гонконгського офісу отримав листа нібито від фінансового директора компанії з британської штабквартири, в якому йшлося, що потрібно провести певний секретний платіж. Працівник спочатку не повірив, справедливо вважаючи, що це спроба фішингу. Тоді зловмисники запросили його підключитись до відеоконференції, у якій брали участь нібито сам фінансовий директор та інші люди з компанії. Насправді всі вони були згенеровані штучним інтелектом. Побачивши, як йому здалося, своїх колег, фінансовий працівник погодився перерахувати суму у 200 млн гонконгських доларів і зробив це 15-ма транзакціями.

Протидія та інновації

Чи можна запобігти таким аферам за допомогою технологій? Тут є два аспекти: розпізнавання підробок і створення рішень, на які підробки не діють. Що стосується другого, то майбутнє IAM, вважає Identity Management Institute, лежить в царині інтелектуальних і адаптивних рішень. Нові технології, такі як системи ідентифікації на базі блокчейну, обіцяють ефективний засіб проти діпфейків. Створюючи незмінювані записи атрибутів особи, блокчейн гарантує, що будь-які маніпуляції з обліковими даними будуть миттєво помічені і зафіксовані.

У цьому відношенні багатообіцяючим напрямком є децентралізовані

системи ідентифікації (DDI), які надають користувачам більше контролю за їхніми обліковими даними. У таких системах дані про користувача не зберігаються на стороні (у банку тощо), а натомість кожен має децентралізований ідентифікатор (DID), який зберігається у блокчейні.

Також перспективними є гібридні продукти IAM, які поєднують контекстову і поведінкову аналітику з моніторингом у реальному часі, забезпечуючи динамічний захист на основі індивідуальних профілів ризику. Ці системи адаптуються до змін в ландшафті загроз і можуть захищати і від традиційних атак, і від тих, що використовують ШІ.

Не «бути в курсі», а дотримуватись протоколу

Про детектори ШІ поговоримо нижче, але Мохаммед Халіл звертає увагу на те, що хоча ринок цих рішень зростає досить непогано (28–42% на рік), це на порядки менше за темпи поширення діпфейків. Відтак вважає, що стратегія захисту, яка базується виключно на придбанні найновіших інструментів, приречена на поразку, а тому потрібні не лише технології, а й надійний процес верифікації, здатний протистояти навіть досконалим підробкам.

Він далі наголошує, посилаючись на різні дослідження, що секрет успішності діпфейкових атак — також не лише технології, а й людська психологія та організаційне недбальство. Існує глибокий і небезпечний розрив між вірою людей у свою здатність розрізняти діпфейки і реальність. Як показують опитування, близько 60% переконані в тому, що зуміють розпізнати згенероване зображення або відео. Між тим у контрольованих експериментах рівень виявлення якісних фейкових відео падав до 24,5%, зображень — до 62%. Стосовно аудіофейків дослідження Флоридського університету показало, що хоча учасники заявляли про рівень викриття у 73%, насправді вони часто бували введені в оману. У ще одному дослідженні лише 0,1% (!) учасників змогли правильно вказати, де фейки, а де справжній контент.

Дослідження бізнесу показують, що компанії загрозу діпфейків серйозно не сприймають. Кожен четвертий керівник зізнається, що не знайомий або погано знайомий з цією технологією, 31% не вірять, що діпфейки збільшують ризик ошуканства у їхній компанії, відтак 80% не мають протоколів реагування на атаки з використанням діпфейків, і понад 50% керівників зізнаються, що їхні співробітники не навчені розпізнаванню діпфейків та поводженню з ними.

Решта, слід так розуміти, навчена, і цей останній пункт вказує на стратегічну помилку в підході до проблеми. Компанії інстинктивно беруться підвищувати обізнаність персоналу і навчати його розрізняти фейки, однак люди просто не перетворяться на фахівців-розслідувачів, особливо під впливом комплексної атаки з використанням соціальної інженерії. В тому, що в Arup видурили $25 млн, винен не працівник, а процеси в організації: політика на кшталт «Усі запити на транзакції на суму понад $10 тис., отримані відеозв'язком або електронною поштою, повинні підтверджуватись голосом по довіреному, зареєстрованому телефонному номеру» зупинила б спробу, якою б довершеною не була підробка.

Прості кроки рятують гроші

Некомерційний фонд OWASP (Open Web Application Security Project), який працює над підвищенням рівня безпеки програмного забезпечення, видав документ із назвою A Guide to Preparing and Responding to Deepfake Events. Його автори так само вважають, що навчання обізнаності про діпфейки неефективне, бо навіює працівникам хибну впевненість, буцімто фейк можна надійно розпізнати за якимись артефактами (неправильні рухи губ чи рук, неприродні паузи в мові тощо). По-друге, для успіху операції діпфейк не обов'язково має бути досконалим, тому що шахраї використовують людську психологію, чинячи тиск на жертву, щоб вона в паніці зробила необдуманий крок. Коли йдеться про успішні атаки з використанням діпфейків, майже завжди це приклад того, як жертву змусили обійти встановлені процедури, а досконалість фейку ролі не грає.

Відтак автори пропонують такі рекомендації щодо навчання:
- не допускати, щоб у працівників склалось хибне враження, ніби вони зуміють розпізнати підробку, а натомість наголошувати на важливості дотримання процедур;
- прагнути, щоб люди пам'ятали, що діпфейки покликані викликати сильні емоції (такі як страх) і вимкнути логіку;
- вчити ставити під сумнів розпорядження начальства, коли воно наказує перейти з автентифікованої платформи на іншу технологію відеоконференцій, посилаючись на «проблемі зі зв'язком», або якщо ці розпорядження надходять незвичними каналами, наприклад, через месенджер;
- наголошувати, що працівники можуть і повинні перевіряти незвичні запити іншими каналами (наприклад, лист — дзвінком на відомий номер, відеодзвінок — поштою і т.д.);
- стандартизувати і поширити практику вимагати верифікації, перш ніж користувач зможе підключитись до конференції;
- використовувати ті ж прийоми, що й на навчаннях по протидії фішингу — робити діпфейки керівників і намагатись сконтактувати зі співробітниками тощо.

Решта рекомендацій OWASP описує процедури, які вже трішки нагадують параною. Зокрема щодо автентифікації автори пропонують впроваджувати принаймні дві практики:
- мати реєстр методів комунікацій, які можна використовувати для додаткової перевірки користувача (наприклад, корпоративний месенджер, додатковий телефон тощо);
- використовувати верифікацію за альтернативним каналом (наприклад, щоб можна було передзвонити на зареєстрований номер і підтвердити особу і запит);
- запровадити щоденні коди (Code of the Day), які передаються через захищений застосунок, по SMS або особисто;
- запровадити для працівників і партнерів питання для перевірки (при цьому заборонити такі як «дівоче прізвище матері» або «кличка домашнього улюбленця»;
- вимагати, щоб начальник або супервізор того, хто дзвонить, підтвердив запит, або дзвонити самому на зареєстрований номер.

Щодо безпеки фінансових операцій рекомендації такі:
- розділити критичні функції таким чином, щоб за транзакцію відповідала більш ніж одна особа — наприклад, авторизувати платіж і обробляти його мають різні люди;
- запровадити подвійну авторизацію — значні транзакції

повинні санкціонувати двоє людей. Тоді кожен платіж здійснюватиметься під контролем іншого працівника;
- запровадити щоденний код для платежів і розкриття чутливої інформації;
- для всіх комунікацій і платежів запровадити багатофакторну автентифікацію;
- ввести верифікацію запитів на платежі двома незалежними каналами зв'язку;
- прописати багатоетапну перевірку для платежів на суму, вищу за певний поріг; тощо.

Документ містить багато інших рекомендацій щодо того, як мають себе поводити, наприклад, працівники гарячої лінії, HR при прийомі на роботу, щодо політик розкриття чутливої інформації, дій у разі виявлення діпфейків і реагування на атаки.

Клин клином

Identity Management Institute зазначає, що штучний інтелект сам по собі забезпечує потужний захист від діпфейків. Створюються алгоритми машинного навчання, які мають виявляти невідповідності і аномалії у штучному контенті. Аналізуючи невідповідності на піксельному рівні і помічаючи, наприклад, неприродні рухи і розсинхронізацію аудіо та відео, ці інструменти зможуть ефективно детектувати підроблені аудіо– та відеозаписи. IAM можуть використовувати ці інструменти для сканування записів на автентичність, а також ці рішення можна використовувати для підтримання безпечності комунікацій і документів у бізнес-процесах.

Наприклад, цими проблемами займаються у Міждисциплінарному центрі безпеки, надійності та довіри (SnT) при Університеті Люксембурга. Як йдеться у статті на їхньому сайті, для виявлення фейків ШІ-моделі навчають методом бінарної класифікації. Наприклад, моделі показують серії зображень, позначених як справжні і підроблені, і вона поступово починає розпізнавати патерни, за якими автентичний контент відрізняється від підробленого.

Проте цей метод має недолік: модель може засвоїти патерни, які не є важливими. Якщо, скажімо, їй згодовували зображення на білому тлі, вона не зможе аналізувати ті, що на чорному. Дослідники виявили, що ефективніше навчати модель виключно на справжніх зображеннях. Якщо дані не відповідають патернам реального зображення, тоді це фейк. Новий підхід робить модель більш стійкою і універсальною.

Іншим викликом є той факт, що з'явився новий клас діпфейків: повністю синтетичні зображення і відео — на відміну від традиційних діпфейків, які все ж базуються на реальному контенті. Це ускладнює виявлення підробок (**рис. 5**). «Більшість детекторів діпфейків добре працюють з заміною і відтворенням облич, але з синтетичними даними вони не працюють. Це рівень генералізації, якого ми прагнемо досягти», — пояснюють дослідники. З цією метою вдалися до багатозадачного навчання, коли ШІ тренують одночасно на кількох аспектах зображення або відео. Замість того, щоб класифікувати об'єкт просто як справжній чи несправжній, модель аналізує різні його аспекти і виявляє місця, які свідчать про підробку.

Рис. 5. На фото дві людини: одна справжня, інші згенерована. Де яка? (Джерело: група SnT при Університеті Люксембурга)

Реальні інструменти для виявлення діпфейків за допомогою ШІ існують і працюють. Приміром, нідерландська компанія Sensity пропонує рішення, яке використовує багатошаровий аналіз (пікселі, метадані, структура файлів, голос). Рішення аналізує відео покадрово і виявляє змінені обличчя, видаючи висновок про справжність чи не справжність, рівень впевненості в результаті і тип маніпуляції. Для виявлення повністю синтетичних облич програма аналізує статистичні і структурні артефакти, такі як асиметрія, нетипові переходи текстур і невідповідність між біологічною геометрією обличчя та ймовірностями, що їх вивчила GAN. Також рішення виявляє клонування і синтез голосу та маніпуляції з аудіозаписами.

Рішення Sensity може бути розгорнуте на сервері з процесорами NVIDIA для досліджень великих масивів даних, на ПК, а також у вигляді USB-пристрою для виїзних груп і для ізольованих (air gapped) серверів. Як стверджує виробник, у 2025 році було виявлено понад 900 тис. інцидентів. Заявлена точність виявлення — до 98%.

Intel FakeCatcher — рішення, яке використовує технологію фотоплетизмографії, тобто фіксації дрібних змін у рухові крові. Система також аналізує рухи очей, які також часто є неприродними на фейкових відео. За даними сайту SOCRadar, FakeCatcher працює на процесорах 3rd Gen Intel Xeon Scalable, підтримує до 72 потоків виявлення одночасно, має заявлену точність 96% в лабораторії і 91% в реальному світі.

McAfee Deepfake Detector працює на ПК з серверами Intel Core Ultra 200V і фокусується на голосі. Програма автоматично сповіщає протягом кількох секунд, якщо користувач переглядає відео, що містить аудіо, згенероване або змінене за допомогою ШІ.

Це лише деякі з детекторів на базі ШІ, які створені для боротьби з діпфейками. Щоправда, експерти висловлюють сумніви щодо їхньої реальної ефективності. Наприклад, за даними Мохаммеда Халіла, на практиці вона падає на 45–50% проти результатів у лабораторії. Додамо: до того ж ці рішення вочевидь не встановиш на будь-якому комп'ютері або телефоні. Тому здорова параноя поки що найкращий наш захист.

Василь ТКАЧЕНКО, **МтБ**