Nie wierz w to, co widzisz

14 lutego 2025

Jeszcze niedawno deepfejki były łatwe do rozpoznania – śmieszne kolaże, źle dopasowane twarze, sztuczna mimika. Potem przyszły płynniejsze animacje i lepiej podmienione głosy, ale wciąż można było dostrzec fałsz. Dziś AI generuje nagrania tak realistyczne, że nawet wprawne oko może się nabrać. Manipulacja przeszła od pojedynczych zdjęć do pełnych sekwencji wideo tworzących rzeczywistość, która nigdy nie istniała. Nie żyjemy już w epoce informacji. Żyjemy w epoce manipulacji.

Wciąż co prawda dominują absurdalne ilustracje wygenerowane przez AI, jak np. dziewczynka z trzema rękami piekąca dom z chałki, bo była sierotą i nie miała gdzie mieszkać, a potem płacząca, bo nikt jej nie pogratulował, albo drwal, który własnoręcznie wyciosał z drewna piękną wannę, po czym usiadł obok niej w głębokim smutku, bo nikt nie docenił jego kunsztu. Na facebookowych grupkach pełno jest serdecznych komentarzy skierowanych do takich wygenerowanych postaci, bo wielu – zwłaszcza tych mniej świadomych technologicznie – nadal nie odróżnia zbyt dobrze fikcji od rzeczywistości.

Ludzie nabierają się nawet na najbardziej pokraczne wygenerowane twory AI, które wrzucane są tylko po to, żeby zebrać jak najwięcej reakcji i podbić zasięgi fanpejdża. A ten, gdy już przyciągnie odpowiednią liczbę klikających w byle co użytkowników, zamienia się w maszynkę do wrzucania podejrzanych linków. Klikniesz? Masz wirusa. Otworzysz? Twoje konto na „Facebooku” właśnie zmieniło właściciela. Zalogujesz się na „super ofertę”? Gratulacje, oszuści właśnie dostali dostęp do twojej bankowości internetowej. I tak oto ci sami ludzie, którzy jeszcze wczoraj ze łzami w oczach gratulowali AI-owemu drwalowi kunsztu, dziś polecają znajomym „pewną inwestycję w kryptowaluty” albo proszą o pilny przelew, bo „utknęli za granicą bez portfela”.

Ale obok tych oczywistych absurdów zaczyna zalewać nas fala hiperrealistycznych wideo generowanych m.in. przez Sorę – narzędzie udostępnione w zeszłym roku przez OpenAI.

Piszę o tym wszystkim, bo od początku roku na mojej facebookowej tablicy zamiast generycznych obrazków biednych dziewczynek zrobionych z chałek i niedocenionych drwali rzeźbiących cuda, pojawia się coraz więcej nagrań, które wyglądają jak zapis prawdziwych wydarzeń. Jeszcze kilka miesięcy temu w sekundę potrafiłbym odróżnić autentyk od AI. Teraz coraz częściej łapię się na tym, że analizuję dłużej i… nie mam pewności.

Od śmiesznych filtrów do hiperrealizmu

Największy niepokój wobec rozwoju AI poczułem, gdy trafiłem ostatnio na wideo przedstawiające gigantyczną lawinę. Sunęła z niewiarygodną siłą, pochłaniając całą dolinę, a obraz wręcz hipnotyzował – ogromne masy śniegu spadały w dół jak biała apokalipsa. Obejrzałem to kilka razy po czym mocniej wciągnąłem się w temat i zacząłem przeglądać inne nagrania lawin na „YouTube” – mam tak, że gdy coś mnie zaciekawi, wsiąkam w to na dobre. Oglądam kolejne filmy, porównuję, szukam dodatkowych informacji, wchłaniając przy tym nową wiedzę.

I wtedy, zupełnym przypadkiem, natrafiłem na to samo nagranie po raz drugi i sprawdziłem sekcję komentarzy, gdzie przewijały się standardowe reakcje: „WOW”, „Niesamowite!”, „Czy wszyscy przeżyli?”.

Jeden użytkownik miał inne zdanie. Napisał coś w stylu: „To wygenerowała SORA od OpenAI. Zwróćcie uwagę na narciarzy”. Początkowo wydawało mi się to przesadą, ale przybliżyłem obraz. A tu… faktycznie, ludzie na stoku niby wyglądali naturalnie, ale ich gesty i reakcje nie do końca pasowały do

wydarzeń. Były minimalnie opóźnione, jakby lekko oderwane od rzeczywistości. Gdyby nie ten jeden komentarz nie spędziłbym kolejnych kilku minut, analizując jeden z wielu wiralowych filmików, który zwyczajnie wyglądał na tyle realistycznie, że dałem się być może po raz pierwszy na coś takiego złapać.

A przecież jeszcze chwilę temu deepfejki wyglądały tak nieporadnie, że w większość przypadków były po prostu śmieszne… Niestety teraz wystarczy parę kliknięć, by w minutę stworzyć hiperrealistyczną symulację zdarzenia, które nigdy się nie wydarzyło.

W 1999 roku „Matrix” zszokował widzów wizją świata, w którym ludzie nie zdają sobie sprawy, że żyją w symulacji. Wtedy to była fantastyka, cyberpunkowa przestroga przed technologią, która przejmuje kontrolę nad percepcją rzeczywistości. Ale dziś nie musimy podłączać się do żadnej gigantycznej symulacji – wystarczy, że scrollujemy „Facebooka”.

Nie potrzebujemy kabli w mózgu ani złowrogiej AI trzymającej nas w kokonie. Sami, dobrowolnie, zanurzamy się w iluzję generowaną przez algorytmy.

Sztuczna inteligencja stopniowo niweluje granicę między tym co autentyczne a tym co fałszywe w sieci. W erze cyfrowej gdzie każdy obraz, nagranie czy tekst może zostać wygenerowany tak po prostu – nie liczy się już co jest prawdziwe – liczy się, w co ludzie są gotowi uwierzyć. A już wcześniej wiele osób wierzyło w rzeczy, które z fałszerstwem AI nie miały jeszcze nic wspólnego – były zwykłymi Paintowymi memami, ale i tak podbijały internetowe społeczności stając się „dowodami” w ich odwiecznych wojenkach.

Tak się składa, że śledzę dyskusje zarówno na tych PiSowskich jak i Platformerskich grupkach. Nie raz widziałem tam grafiki takie jak np. nieudolnie wklejony na tle Kremla Kaczyński świętujący urodziny Putina z kieliszkiem szampana w dłoni.

Mimo że oświetlenie twarzy kompletnie nie pasowało do reszty

sceny i jego twarz w porównaniu do reszty „dowodu” była rozpiskelizowana to nie przeszkadzało to kodziarskiej grupce w szerzeniu tej „sensacji”. Pod zdjęciem wrzało: „KACZYŃSKI = KREML, TO JEST AGENTURA, ROZPOWSZECHNIAĆ PÓKI NIE USUNĄ!!!!!”, „Hańba! Tak sprzedali Polskę, Putin zadowolony! A suweren dalej klęczy przed Kaczafim”, „Ile jeszcze dowodów potrzeba na to, że PiS to ruska partia???”. Oczywiście w sekcji komentarzy prędzej można było znaleźć kłótnie o przecinki niż jakąkolwiek refleksję nad jakością zdjęcia.

Z drugiej strony, po pisowskiej stronie internetu krążyło np. równie absurdalne zdjęcie Tuska w mundurze Wehrmachtu, stojącego ramię w ramię z Angelą Merkel i krzyczącego – a jakżeby inaczej – „FÜR DEUTSCHLAND”.

Nie będę już opisywał jak tragicznie było to wykonane. Ale dla tych, którzy od lat widzieli w Tusku „niemieckiego agenta”, było to potwierdzenie wszystkiego w co i tak już wierzyli. „NIEMIECKA ŚWINIA SPRZEDAŁA NAS ZA 30 SREBRNIKÓW!!!!!”, „Hańba!! Kiedy TRYBUNAŁ STANU??? Merkel TRZYMA GO na krótkiej smyczy!!!”, „Tusku ty Judaszu BĘDZIESZ SIEDZIAŁ”. W sekcji komentarzy standardowo pełno było błędów ortograficznych, a gdy ktoś sugerował, że to fotomontaż, natychmiast odpowiadano mu: „LEWAKU, ZaMKNIJ R*J NIC JUŻ WAS NIE URATUJE!!!!!”, „Lewactwo POd sąd i do kamieniołomów!!!” i tak dalej bo dla niektórych liczy się tylko to, w co ktoś chce wierzyć.

Jednak odchodząc już od przykładów politycznych. Do najbardziej absurdalnych rzeczy zaobserowanych podczas moich internetowych wojaży można zaliczyć chyba zdjęcie Jana Pawła II przerobione przez najprostszy filtr Snapchata, który nadał mu długie blond włosy i lekki makijaż. Ktoś wrzucił je na grupę randkową z podpisem „Poznam Pana/Panią”, a w komentarzach posypały się wiadomości od adoratorów. „Śliczna, kobieta z klasą w oczach”, „Ale cudna, zapraszam na kawkę do siebie, umiem gotować”… „To się nazywa prawdziwa dama, nie jakieś dzisiejsze d***ki” i tak dalej i tak dalej.

A to wszystko działo się jeszcze zanim AI weszło na poziom generowania hiperrealistycznych filmów i zdjęć. Jeśli ludzie potrafili bezrefleksyjnie przyjąć tak nieudolne fotomontaże jako fakty, to co będzie dalej?

Zbyt niebezpieczne, by pozostawić to bez kontroli

Parę lat temu stworzenie realistycznego materiału wideo było kosztownym i czasochłonnym przedsięwzięciem. W Hollywood całe zespoły grafików, programistów i animatorów tygodniami dopracowywały każdą pojedynczą klatkę efektów specjalnych, a mimo wciąż potrafiliśmy dostrzec granicę między sztucznością a rzeczywistością. Dziś ta granica zanika, bo wystarczy parę kliknięć, aby wygenerować dowolny obraz, dowolną osobę i dowolne słowa, które nigdy nie zostały wypowiedziane.

W 2024 roku OpenAI zaprezentowało Sorę – model sztucznej inteligencji, który potrafi tworzyć pełnowymiarowe, hiperrealistyczne filmy na podstawie krótkiego opisu tekstowego. Wpisujesz: „Mężczyzna idzie ulicą w deszczu, ubrany w płaszcz, jego twarz odbija się w mokrym asfalcie” – i po chwili masz gotowe nagranie, które wygląda jak rzeczywiste. Nie jest to animacja, nie jest to montaż – to zupełnie nowa rzeczywistość wykreowana od zera przez sztuczną inteligencję.

To oznacza koniec monopolu wielkich studiów filmowych, które jako jedyne mogły sobie pozwolić na efekty specjalne za miliony dolarów, ale też początek epoki, w której prawda stanie się opcjonalna. Możliwości tej technologii są może dla niektórych fascynujące, ale dla mnie mimo wszystkich pozytywów są zwyczajnie przerażające.

Sora jest naprawdę niezła jeśli chodzi o produkcje nieistniejących filmów, ale jeszcze bardziej obawiam się

chińskiego OmniHuman-1 – rozwijanego przez ByteDance, właściciela TikToka.

W ostatnio zademonstrowanym narzędziu (Póki co jeszcze niedostępnym publicznie) wystarczy jedno zdjęcie, by system „ożywił” dowolną osobę, nadając jej ekspresję i ruchy, które wyglądają niezwykle naturalnie. Głos można również bardzo szybko wygenerować za pomocą innych dostępnych AI na podstawie dowolnego krótkiego nagrania danej osoby. Największą furorę w sieci zrobiło 23-sekundowe nagranie „ożywionego” Alberta Einsteina, które wprawiło ekspertów w osłupienie jako jeden z najbardziej realistycznych deepfejków w historii.

Henry Ajder, ekspert zajmujący się analizą deepfejków, ostrzega: „Jeszcze niedawno stworzenie realistycznego deepfake’a wymagało setek zdjęć i ogromnych zasobów. Teraz wystarczy jedno zdjęcie, by stworzyć wideo, które wygląda jak autentyczne”.

Możliwości tej technologii są szerokie – od edukacji, przez rekonstrukcję historyczną, aż po przemysł rozrywkowy i marketing. Ale równie dobrze może ona posłużyć do tworzenia fałszywych dowodów w sprawach sądowych, kompromitujących nagrań, szantaży czy manipulowania opinią publiczną na niespotykaną dotąd skalę.

Sora i OmniHuman-1 to przełom, który nigdy nie powinien trafić w ręce kogokolwiek bez specjalnej kontroli i odpowiednich regulacji.

Nowy arsenał do oszustw i szantaży

Przestępcy musieli wcześniej stosować naprawdę nieudolne metody w porównaniu z tym co mają obecnie – słabej jakości

rozmowy telefoniczne czy masowe wiadomości e-mail o wujku z USA, który chce ci przekazać milionowy spadek. W epoce AI wystarczy nagranie głosu, kilka filmików, a może nawet jedno zdjęcie z Instagrama, a sztuczna inteligencja stworzy w pełni realistyczną osobę, która poprosi o przelew albo zmanipuluje człowieka do podjęcia decyzji, której nigdy by nie podjął.

W 2024 roku do pracowników banku w Hongkongu zadzwonił na wideokonferencji ich własny dyrektor. Znajomy ton głosu, naturalna mimika, ten sam sposób mówienia. Nic nie budziło najmniejszej wątpliwości, więc pieniądze – 35 milionów dolarów – zostały przelane na wskazane konto. Dopiero później okazało się, że dyrektor nigdy nie zlecał tego przelewu, a jego postać została całkowicie wygenerowana przez AI.

Aleksandra Przegalińska zwraca uwagę, że współczesna sztuczna inteligencja rozwija się nie tylko w zakresie realistycznej imitacji głosu i obrazu, ale również w umiejętności analizowania kontekstu komunikacji i dostosowywania się do sposobu wyrażania myśli przez konkretne osoby. W książce Sztuczna Inteligencja. Nieludzka, arcyludzka opisuje, jak zaawansowane modele AI mogą odwzorowywać styl wypowiedzi, a nawet charakterystyczne reakcje użytkowników, co znacząco utrudnia wykrywanie fałszerstw. Przegalińska podkreśla, że tego rodzaju technologia nie ogranicza się już do samego podrabiania wizerunku – pozwala na symulację zachowań i sposobu myślenia danej osoby w stopniu, który sprawia, że granica między prawdą a fikcją staje się coraz mniej wyraźna. W efekcie manipulacje dokonywane przy użyciu AI mogą być bardziej przekonujące i trudniejsze do wykrycia niż tradycyjne metody oszustwa.

Oczywiście nie trzeba być dyrektorem banku, by paść ofiarą deepfejkowego szfindla. Oszuści coraz częściej używają ich w internetowych wyłudzeniach miłosnych. Samotne osoby wierzą, że nawiązały kontakt z atrakcyjną kobietą lub mężczyzną (np. bogatym amerykańskim pułkownikiem), podczas gdy tak naprawdę rozmawiają z cyfrową iluzją.

Historia 53-letniej Francuzki, która uwierzyła, że koresponduje z Bradem Pittem, brzmi jak żart, ale dla niej skończyła się dramatem. Przez kilka miesięcy otrzymywała sfabrykowane filmy i wiadomości głosowe, które wyglądały i brzmiały dokładnie tak, jakby należały do aktora. Oszust przekonał ją, że jest poważnie chory, a ona, poruszona historią ukochanego, przelała mu ponad 3 miliony złotych i rozwiodła się z mężem. Dopiero kiedy zobaczyła prawdziwego Pitta w mediach, zorientowała się, że padła ofiarą doskonałej mistyfikacji.

Scamy miłosne to tylko część góry lodowej. Deepfejk coraz częściej jest używany również do szantażu. Nie trzeba już prawdziwych kompromitujących materiałów – wystarczą czyjeś publicznie dostępne zdjęcia z internetu, by stworzyć zdjęcia lub film, na którym ofiara rzekomo bierze udział w czymś, czego nigdy nie zrobiła. Przestępcy żądają pieniędzy, grożąc publikacją, a wiele osób, zamiast ryzykować społeczne konsekwencje, po prostu płaci.

Co gorsza deepfejki podważają fundamenty prawne i społeczne. Kiedyś nagranie audio lub wideo było dowodem nie do zbicia – teraz to broń, którą można stworzyć w minutę, a potem równie łatwo podważyć. W Wielkiej Brytanii facet sfabrykował dowody na zdradę żony, żeby lepiej na tym wyjść w rozwodzie. W USA kobieta przez kilka miesięcy była oskarżona o groźby karalne, bo ktoś podmienił jej głos na nagraniu. Dopiero eksperci wykazali, że nigdy tych słów nie powiedziała.

I to nie są jakieś odległe historie z gazet – to problem, który za chwilę może dotknąć każdego. Wystarczy jeden rozwód, jedna kłótnia, jeden konflikt w pracy. Kiedyś, jeśli ktoś pokazał nagranie zdrady czy szantażu, sprawa była jasna. Dziś? „To nie ja, to deepfejk” – i nagle nie oszust musi bronić swojej niewinności, tylko ofiara musi udowodnić, że dowód nie jest fałszywką.

Mustafa Suleyman w książce Nadchodząca fala zwraca uwagę, że

deepfejki tworzą podwójne zagrożenie – nie tylko fałszują rzeczywistość, ale też podkopują wiarygodność tego, co jest prawdziwe.

„Nadchodzi era, w której ludzie nie tylko będą regularnie oszukiwani przez syntetyczne treści, ale także przestaną ufać autentycznym materiałom. Zamiast szukać prawdy, zaczną wybierać tę wersję rzeczywistości, która im najbardziej odpowiada.”

Twory AI jako broń polityczna

No i przechodzimy do najbardziej niebezpiecznego dla naszej kochanej demokracji zastosowania tego dynamicznego rozwoju technologii.

Technologia generatywnej sztucznej inteligencji rozwinęła się do tego stopnia, że fałszywe nagrania polityków, zmanipulowane oświadczenia czy sztucznie wykreowane skandale są niemal nie do odróżnienia od prawdy. To nie scenariusz dystopijnego filmu, ale nasza rzeczywistość.

Jeden ze znanych przypadków wykorzystania deepfejków w kampanii politycznej miał miejsce w Indiach, gdzie partia BJP wykorzystała technologię, aby jej lider Manoj Tiwari, „przemawiał” do wyborców w różnych dialektach, mimo że w rzeczywistości nie znał tych języków. Nagranie zostało zmienione tak, by jego usta poruszały się synchronicznie z generowanym dźwiękiem tworząc iluzję, że rzeczywiście wypowiada słowa w lokalnych językach. Oficjalnie miało to na celu dotarcie do szerszego elektoratu, jednak przypadek ten wywołał obawy o przyszłe nadużycia tej technologii. Jeśli deepfejk można wykorzystać do tłumaczenia polityków, nic nie stoi na przeszkodzie, by zmieniać ich wypowiedzi w sposób bardziej subtelny – fabrykować obietnice wyborcze, osłabiać przeciwników czy kreować nieistniejące skandale.

Przekonali się o tym Mołdawianie, gdy w grudniu 2023 roku

prezydent Maia Sandu padła ofiarą postępu technologicznego. Fałszywy materiał przedstawiał ją jako osobę ironizującą na temat poziomu życia Mołdawian i sugerującą, że jej polityka jest finansowana przez George'a Sorosa i USA – klasyczny motyw rosyjskiej dezinformacji. Nagranie pojawiło się tuż przed jej oficjalnym noworocznym przemówieniem i było szeroko dystrybuowane w rosyjskojęzycznych kanałach na Telegramie, gdzie takie ataki informacyjne są obecnie standardową bronią Kremla. To nie był odosobniony przypadek – zaledwie miesiąc wcześniej przed wyborami lokalnymi w sieci pojawiły się kolejne deepfejki, w których Sandu rzekomo ogłaszała swoją rezygnację i namawiała do głosowania na prorosyjską partię.

Podobna technika została użyta w USA, gdzie w 2023 roku rozpowszechniano fałszywe nagranie Joe Bidena, który miał wprowadzać obowiązkowy pobór do wojska. Chociaż film został szybko zdementowany, zdążył wywołać panikę i falę oburzenia wśród wyborców. Wcześniej w 2019 roku zmanipulowane nagranie Nancy Pelosi przedstawiające ją jako osobę mamroczącą i niezdolną do prowadzenia debaty było szeroko rozpowszechniane przez prawicowe media i stało się pretekstem do podważania jej kompetencji.

Deepfejki są już codziennym narzędziem wojny hybrydowej. W 2022 roku na Telegramie pojawiło się nagranie prezydenta Ukrainy Wołodymyra Zełenskiego, który rzekomo ogłaszał kapitulację wobec Rosji. Choć zostało to szybko zdementowane, zdążyło wywołać falę dezinformacji i paniki. Co gorsza, deepfejk trafił nie tylko do internetu – pojawił się również w zhakowanych ukraińskich kanałach telewizyjnych, co nadało mu dodatkową wiarygodność i zwiększyło chaos informacyjny. W podobny sposób rosyjskie farmy trolli rozpowszechniały zmanipulowane „wywiady" z fikcyjnymi ekspertami, którzy analizowali wojnę na Ukrainie, szerząc kremlowską propagandę.

Problem z deepfejkami, ponownie przypominam, nie polega jedynie na tym, że mogą skutecznie fałszować rzeczywistość. Prawdziwą tragedią jest to, że mogą podważać również prawdziwe

wydarzenia. Jeśli polityk zostaje przyłapany na kompromitujących słowach albo braniu łapówki to zawsze może powiedzieć, że to fałszywka. I w ten sposób krok po kroku deepfejki burzą fundamenty demokracji i publicznego zaufania. Bo jeśli każde nagranie można podważyć jako zmanipulowane, to w końcu ludzie przestają wierzyć we wszystko. I wtedy zwyciężają ci, którzy najlepiej zarządzają kłamstwem.

Zmęczenie dezinformacją

Ludzki mózg nie został zaprogramowany na świat, w którym prawda i fałsz zlewają się w jedno. Lawina deepfejków, generowanych treści i zmanipulowanych materiałów doprowadza do informacyjnego przesycenia, a zaufanie do rzeczywistości zaczyna pękać jak ściany w nowym mieszkaniu od chciwego dewelopera, który ewidentnie oszczędzał na materiałach. W efekcie społeczeństwo przechodzi w tryb „wszystko mi jedno" – ludzie przestają analizować, co jest prawdziwe, a co spreparowane, bo zwyczajnie nie mają już na to siły.

I to nie jest żadna abstrakcyjna teoria. Badanie przeprowadzone w 2024 roku przez University of Cambridge i MIT Media Lab pokazało, że 40 proc. uczestników miało problem z odróżnieniem autentycznych nagrań od deepfejków, a 25 proc. uznało, że w sumie to i tak wszystko jest zmanipulowane – więc po co w ogóle się przejmować? Naukowcy nazwali to „zmęczeniem poznawczym dezinformacją", czyli stanem, w którym zamiast próbować cokolwiek zweryfikować, ludzie machają ręką i traktują każdą treść jako potencjalne kłamstwo. Bo skoro wszystko może być fejkiem, to równie dobrze nic nie musi być prawdziwe.

W praktyce oznacza to, że im więcej deepfejków zalewa internet, tym mniej ludzi w ogóle zawraca sobie głowę odróżnianiem prawdy od fałszu. Skandale polityczne? Fejk. Przecieki z rządu? Ustawka. Nagranie przestępstwa? Kolejna prowokacja. Każdy materiał może być zmanipulowany, więc

zamiast dociekać, czy coś jest prawdziwe, społeczeństwo po prostu wzrusza ramionami i idzie dalej. Zaczynamy funkcjonować w rzeczywistości, gdzie reakcją na każdą informację jest bezrefleksyjne „może tak, może nie, co za różnica". Dezinformacja nie musi już niczego udowadniać – wystarczy, że zasieje wystarczająco dużo wątpliwości, by ludzie przestali przejmować się czymkolwiek.

„Minęło, nie minęło. Prawda, nieprawda. Wszystko jedno". – tak właśnie działa mechanizm kontroli przez chaos, gdzie nie trzeba nawet fałszować rzeczywistości… wystarczy ją rozmyć.

Fałsz, na którym można zarobić

Deepfejki to technologiczna hydra – odrąbiesz jedną głowę, a na jej miejscu wyrastają dwie kolejne, jeszcze lepiej przygotowane na detekcję. Sztuczna inteligencja nie tylko uczy się omijać systemy wykrywania, ale robi to szybciej, niż ktokolwiek zdąży opracować skuteczną metodę obrony. Walka z nimi toczy się na trzech poziomach: technologii detekcji, regulacji prawnych i edukacji społecznej. I choć edukacja faktycznie działa jeśli jest efektywnie wdrażana – bo ludzie stają się coraz bardziej świadomi, coraz więcej osób rozpoznaje manipulację, coraz rzadziej dają się nabierać na najbardziej oczywiste fejki – to gdy przybywa coraz więcej fotorealistycznych syntetycznych treści, sama świadomość już nie wystarcza.

Systemy wykrywania deepfejków początkowo skupiały się na wychwytywaniu błędów – sztucznych mrugnięć, źle padających cieni, dziwnego zniekształcenia twarzy. Przez chwilę to działało, ale AI jest jak cwany oszust i szybko nauczyła się maskować swoje ślady. Badania MIT Media Lab z 2024 roku wykazały, że skuteczność tych metod wynosi maksymalnie 70 proc., czyli co trzeci deepfejk przechodzi jako autentyk. W badaniu Facebook DeepFake Detection Challenge (DFDC) przeprowadzonym 4 lata wcześniej najlepsze algorytmy osiągały

maks 75 proc. skuteczności. Wniosek? Możemy budować coraz lepsze narzędzia do wykrywania ingerencji sztucznej inteligencji, ale AI rozwija się szybciej niż te narzędzia. Im bardziej próbujemy ją przechytrzyć, tym sprawniej dostosowuje się do nowych warunków.

Big Tech udaje, że walczy ze złym wykorzystywaniem AI, ale w rzeczywistości sam nakręca ten chaos. Dajmy na przykład OpenAI – z jednej strony wypuszcza Sorę, czyli model AI zdolny do generowania hiperrealistycznych filmów, a z drugiej oczywiście pięknie zapewnia, że pracuje nad mechanizmami „wykrywania deepfejków". Brzmi ładnie, ale nic za tym nie idzie i coraz to lepsze deepfejki latają po sieci. Firma jest trochę jak landlord, który wynajmuje mieszkanie z grzybem na ścianie, a potem planuje ci sprzedać odgrzybiacz.

OpenAI wymaga, żeby filmy generowane przez Sorę miały oznaczenie… o ile nie zapłacisz za wersję Pro, gdzie możesz pobierać je bez widocznego znaku wodnego. A co potem? OpenAI już się tym nie interesuje. Wystarczy pobrać wideo, wrzucić je gdziekolwiek i nagle nikt nie wie, czy to prawdziwe nagranie, czy generatywna iluzja. Oficjalne stanowisko? „To nie nasza wina".

Meta to chyba najbardziej podręcznikowy przykład Big Techowego zepsucia. W 2019 roku firma uruchomiła Facebook DeepFake Detection Challenge (DFDC) – wielką inicjatywę, w którą wpakowano miliony dolarów, by znaleźć najlepsze algorytmy wykrywające deepfejki. Po latach badań i szumnych obietnic… nic. Żaden z tych systemów nie został wdrożony na szeroką skalę. Meta tłumaczyło to „technicznymi ograniczeniami", ale rzeczywistość jest prosta jak obsługa cepa – walka z deepfejkami się im zwyczajnie nie opłaca. Przecież to właśnie treści wygenerowane przez generatywną sztuczną inteligencję przyciągają ludzi, wywołują reakcje i podbijają zaangażowanie. W końcu w internecie liczy się jedno… żebyś scrollował dalej, oglądał reklamy i trenował algorytm, na którym później ktoś tam na górze sobie zarabia.

Właściciele „Facebooka” co prawda wprowadzili etykiety ostrzegawcze przed treściami wygenerowanymi przez AI, ale są one opcjonalne dla osoby postującej i 95 proc. użytkowników i tak w nie nie klika, więc stosowane przez platformę „ostrzeżenia” to tylko atrapa. Służą jedynie do tego, żeby Meta mogła udawać, że problem jest pod kontrolą, a w praktyce nie robi z nim tyle co nic.

Podobnie wygląda kwestia moderacji. „Facebook” oficjalnie „zakazuje” deepfejków w polityce, ale ma jeden malutki wyjątek: jeśli fejk zostanie oznaczony jako „satyryczny”, to hulaj dusza, piekła nie ma. Ktoś wrzuca zmanipulowane wideo z podpisem „hehe, beka, to tylko żart”, „Facebook” przyklepuje, bo przecież satyra jest dozwolona. A potem wystarczy, że ktoś skopiuje filmik, usunie podpis i wrzuci jeszcze raz – i bum, nagle to już nie żaden „żart”, tylko „szokujące ujawnienie prawdy”. Fejk zaczyna krążyć po sieci, a ludzie udostępniają bez zastanowienia.

To tak, jakby właściciel baru rozlewał ludziom wódę, a potem udawał, że nie ma nic wspólnego z burdą na zapleczu, bo przecież „on tylko wynajmuje stoliki”. „Facebook” dobrze wie, jak to działa, ale nie zamierza nic z tym zrobić, bo hajs się zgadza.

A jeśli myśleliście, że Meta nie może bardziej podkopać własnej wiarygodności w walce ze zbrodniczym AI, to w tym roku zlikwidowała współpracę ze znienawidzonymi, przez nowego przyjaciela Zuckerberga, niezależnymi weryfikatorami treści i zastąpiła ich „Notatkami Społeczności” – systemem, w którym użytkownicy… głosują, co jest prawdą. Nie liczy się analiza ekspertów, nie liczą się fakty, tylko to, która wersja rzeczywistości ma więcej lajków.

Jak to działa? Notatki Społeczności pozwalają użytkownikom dodawać kontekst do postów i oceniać, czy dana informacja jest rzetelna. Jeśli odpowiednia liczba osób z „różnych baniek informacyjnych” uzna, że komentarz jest pomocny, zostaje on

przypięty jako dodatkowe wyjaśnienie. W teorii brzmi to jak demokratyzacja fact-checkingu, w praktyce oznacza, że to, co uznawane jest za prawdę, zależy od opinii użytkowników – a niekoniecznie od rzeczywistości.

Przykład? Ktoś wrzuca post o tym, że Bill Gates planuje zaczipować ludzi, PizzaGate było prawdziwe, a Ziemia jest płaska, bo „NASA to wielki przekręt". Jeśli odpowiednia grupa osób zadecyduje, że brzmi to sensownie, notatka pod postem może nawet nie podważyć tych rewelacji. Z kolei ktoś, kto napisze, że jednak szczepionki nie sterują ludzkim DNA i Elon Musk nie jest reptilianinem, może zostać uznany za mało wiarygodne źródło – jeśli trafi na złe grono oceniających.

Obecnie są dostępne tylko w USA i kilku innych krajach anglojęzycznych, więc nasz internet (przynajmniej na razie) nie jest nimi objęty głównie dlatego, że w Unii Europejskiej obowiązują bardziej rygorystyczne przepisy dotyczące dezinformacji, co może opóźnić ich wdrożenie.

Chociaż warto wspomnieć, że dla nas weryfikacja zgłoszonego posta nie wygląda wcale jakoś lepiej niż wspomniane Notatki Społeczności. Możesz zgłosić coś jako dezinformację, a potem czekać na odpowiedź w stylu: „Dziękujemy za zgłoszenie. Po dokładnym przeanalizowaniu stwierdziliśmy, że treść nie narusza naszych standardów społeczności". Czyli fejk może dalej krążyć, bo algorytym Mety uznał, że jej to nie obchodzi (szczególnie gdy na tym zarabia).

Od długiego czasu widuję na „Facebooku" i „Instagramie" deepfejkowe reklamy – filmy, w których prawdziwi znani polscy aktorzy, gwiazdy, politycy mówią zmyślone rzeczy, bo AI podmieniło im dźwięk. Reklamują cudowne leki, rzekomo finansowane przez Ministerstwo Zdrowia, które wyleczą każdą chorobę, jaką kiedykolwiek wymyślił jakiś naciągacz.

Zgłaszam te reklamy regularnie. I co? Nic. Jak były, tak są. „Facebook" bierze kasę, oszukane osoby klikają i tracą

pieniądze, ktoś zgłasza oszustwo, ale algorytm uznaje, że nie narusza regulaminu. A jeśli nawet reklama po jakimś czasie znika, to natychmiast pojawiają się nowa – ten sam fejk tylko reklamowany przez inną znaną osobowość.

Więc czy Meta faktycznie walczy z deepfejkami? Oczywiście, że nie. Meta je hoduje, karmi i inkasuje na nich zyski, a potem udaje, że walczy.

Regulacje, regulacje, regulacje

Unia Europejska próbowała postawić tamę dla deepfejków, ale AI Act ogranicza się głównie do nakazu oznaczania syntetycznych treści i ich identyfikacji. Tylko co z tego, skoro nikt nie ma realnych narzędzi, żeby te przepisy egzekwować? Technologia pędzi szybciej niż unijna biurokracja, więc gdy regulacje wchodzą w życie, deepfejki są już o krok do przodu, a AI nauczyło się je omijać.

Do tego dochodzi jeszcze jeden problem: prawo działa tylko w UE, a internet granic nie uznaje. Wystarczy, że deepfejk powstanie na serwerach w krajach, które mają regulacje w głębokim poważaniu – i problem znika… ale tylko z unijnych statystyk. Nawet gdyby UE hipotetycznie zakazała Sorze działania na swoim terenie (choć na razie OpenAI samo postanowiło jej tu nie udostępniać), wystarczy włączyć VPN na USA i jazda – można generować deepfejki bez żadnych ograniczeń.

Stany Zjednoczone podeszły do problemu deepfejków w klasycznym amerykańskim stylu: niech każdy stan radzi sobie sam. Na poziomie federalnym nie istnieją żadne przepisy dotyczące syntetycznych treści. Tylko kilka stanów – Kalifornia, Teksas i Nowy Jork – próbowało coś z tym zrobić, regulując polityczne deepfejki i oszustwa seksualne, ale brak jednolitego prawa sprawia, że te przepisy są dziurawe jak program Konfederacji. Więc w teorii mamy jakieś regulacje, a w praktyce? Klasyczne „nikt nic nie wie”.

Chiny jako jeden z nielicznych krajów postanowiły wziąć deepfejki za mordę i wprowadziły obowiązek ich oznaczania. Brzmi rozsądnie? Może i tak, ale haczyk jest oczywisty – przepisy dotyczą głównie deepfejków politycznych i wszystkiego, co mogłoby zaburzyć „stabilność społeczną”. A cała reszta? Syntetyczne reklamy, manipulacje w e-commerce, AI-owe klony celebrytów wciskające podejrzane produkty? To już inna bajka. Nie chodzi więc o to, żeby deepfejki zniknęły, tylko o to, żeby były pod właściwą kontrolą.

Big Tech nie walczy z deepfejkami – on je hoduje. Bo dla nich sztuczna inteligencja to nie problem, tylko złote jajo, które generuje ruch, podkręca monetyzację i staje się narzędziem politycznych rozgrywek. Dlatego problem deepfejków nie zostanie rozwiązany, dopóki kontrolę nad nimi sprawują te same firmy, które zarabiają na ich istnieniu. I dopóki rządy zamiast realnych działań będą wprowadzać regulacje, które można obejść w trzy sekundy.

Autorstwo: Julian Mordarski
Źródło: Trybuna.info