

Autor Max Paradox

SERIA BRUTALNA PRAWDA

Brutalna prawda o deepfake'ach

Kiedy własnym oczom nie można już ufać



SPIS TREŚCI

INTRO.....	3
Rozdział 1 - Widziałem to na własne oczy.....	13
Rozdział 2 - Głos mamy w obcym telefonie.....	22
Rozdział 3 - To przecież naprawdę jest on.....	31
Rozdział 4 - Sprostowanie przyszło, ale nikt już nie patrzył.....	40
Rozdział 5 - Kiedy dowód przestaje kończyć rozmowę.....	49
Rozdział 6 - Twoja twarz już nie potrzebuje ciebie.....	57
Rozdział 7 - Każdy może dziś powiedzieć, że to nie on.....	65
Rozdział 8 - Film był fałszywy, wstyd został.....	73
Rozdział 9 - Nawet rodzinne nagranie zaczyna potrzebować certyfikatu.....	81
Rozdział 10 - Skoro wszyscy to widzieli, coś musi w tym być	89
Rozdział 11 - Kiedy przeprosiny nie wystarczają, bo niczego nie zrobiłeś.....	97
Rozdział 12 - Prawda też może wyglądać podejrzanie.....	106
Rozdział 13 - Kiedy twoje własne wspomnienie zaczyna przegrywać z filmem.....	114
Rozdział 14 - Kiedy kompromitacja staje się usługą na zamówienie.....	122
Rozdział 15 - Kiedy nie wiesz już, czy człowiek naprawdę do ciebie mówi.....	129
Rozdział 16 - Kiedy twoje milczenie też można wygenerować	136

Rozdział 17 - Kiedy człowiek zaczyna bać się własnego archiwum.....	143
Rozdział 18 - Kiedy cynizm zaczyna wyglądać jak inteligencja.....	151
Rozdział 19 - Kiedy prawda potrzebuje rzecznika prasowego	159
Rozdział 20 - Świat się wydarzył, ale proszę to udowodnić	167

INTRO

Telefon leży na stole pomiędzy kubkiem z kawą a talerzem, na którym stygnie śniadanie, ale nikt już nie patrzy na jedzenie. Na ekranie znany polityk mówi coś tak absurdalnego, że pierwszą reakcją nie jest nawet oburzenie, tylko krótkie, triumfalne: "Wiedziałem". Ktoś przesuwając telefon w stronę drugiej osoby, żeby również mogła zobaczyć dowód na to, że człowiek, którego od dawna uważała za idiotę, właśnie publicznie potwierdził wszystkie wcześniejsze podejrzenia. Głos się zgadza, twarz się zgadza, sposób poruszania ust wygląda przekonująco, w tle znajduje się sala konferencyjna, a pod nagraniem zdążyło już pojawić się kilkanaście tysięcy komentarzy. Dopiero ktoś trzeci, najwyraźniej pozbawiony potrzeby natychmiastowego zwycięstwa, wpisuje kilka słów w wyszukiwarkę i odkrywa, że nagranie nigdy się nie wydarzyło. Polityk nie powiedział tych słów, konferencja była prawdziwa, obraz częściowo również, głos został wygenerowany, ruch ust poprawiony, a najbardziej autentyczna w całym materiale była reakcja widzów. Nikt nie został oszukany wyłącznie przez technologię. Technologia jedynie podała gotowy dowód ludziom, którzy bardzo chcieli, żeby był prawdziwy.

Deepfake kojarzył się przez pewien czas z internetową ciekawostką, technologiczną sztuczką i filmikami, w których twarz aktora przyklejano do ciała kogoś innego z efektem wystarczająco niedoskonałym, żeby można było się pośmiać. Było w tym coś niewinnego, ponieważ fałsz sam zdradzał własną obecność. Skóra wyglądała zbyt gładko, oczy zachowywały się dziwnie, usta nie trafiały dokładnie w słowa, a całość przypominała cyfrowego manekina, który bardzo stara się przekonać widza, że ma puls. Ten etap był

komfortowy, bo pozwalał zachować poczucie przewagi. Człowiek siedzący przed ekranem mógł uznać, że technologia może być sprytna, ale on przecież widzi różnicę. Rozpoznaje sztuczność. Zauważa szczegóły. Nie da się złapać na tani numer. Problem zaczął się wtedy, gdy numer przestał być tani, a różnice stały się wystarczająco drobne, żeby przestały przeszkadzać komuś, kto i tak ogląda materiał przez kilka sekund na małym ekranie telefonu między jedną wiadomością a drugą.

Nie trzeba nawet tworzyć doskonałego fałszu, żeby osiągnąć pożądaną efekt, ponieważ odbiorca rzadko zachowuje się jak ekspert sądowy analizujący materiał klatka po klatce. Ogląda nagranie w autobusie, podczas przerwy w pracy, w łóżku przed snem albo w kolejce do kasy. Ma uwagę podzieloną między film, powiadomienia, własne myśli i lekkie poczucie, że powinien robić coś innego. W takich warunkach wystarczy, że obraz wygląda prawdopodobnie, głos brzmi znajomo, a treść zgadza się z tym, czego odbiorca już wcześniej spodziewał się po pokazanej osobie. To właśnie ten ostatni element okazuje się najbardziej użyteczny. Deepfake nie musi przekonywać człowieka do rzeczy całkowicie obcej. Znacznie łatwiej podsunąć mu wersję świata, którą już częściowo nosi w głowie, a następnie dodać do niej twarz, głos i ruch ust. Technologia nie musi wtedy budować wiary od zera. Wystarczy, że dostarczy brakujący rekwizyt.

Wyobraź sobie wiadomość głosową od własnego dziecka. Ten sam ton, ten sam sposób skracania zdań, znajome "mamo" albo "tato", którego nie da się pomylić z nikim innym, a potem informacja, że wydarzył się wypadek, telefon zaraz się rozładuje i trzeba natychmiast przestać pieniądze. Nie ma czasu na analizę widma dźwięku ani poszukiwanie cyfrowych

artefaktów. Jest ciało, które reaguje przed rozumem. Serce przyspiesza, dłonie robią się zimne, myśl zwęża się do jednego celu: pomóc. W takim momencie pytanie o autentyczność może zostać odebrane przez własny umysł niemal jak forma okrucieństwa, bo skoro głos brzmi prawdziwie, to każda sekunda sprawdzania wydaje się sekundą straconą. Deepfake wchodzi więc nie tylko w obszar percepcji, ale również w miejsca, w których człowiek jest najbardziej podatny na skrócenie procesu myślenia. Strach, miłość, gniew, pożądanie, zazdrość, poczucie zagrożenia i potrzeba natychmiastowego działania są znacznie lepszymi współnikami manipulacji niż najbardziej zaawansowany algorytm.

Można oczywiście uspokajać się przekonaniem, że wystarczy być rozsądnym, sprawdzać źródła i nie wierzyć we wszystko, co pojawia się w internecie. Brzmi to bardzo elegancko, szczególnie wypowiedane po fakcie, kiedy wiadomo już, które nagranie było fałszywe, a które prawdziwe. Rozsądek ma jednak nieprzyjemną właściwość: działa najlepiej wtedy, gdy człowiek nie jest emocjonalnie zaangażowany. Łatwo dostrzec manipulację w filmie dotyczącym obcej osoby, której los jest nam obojętny. Znacznie trudniej zachować ten sam chłód, kiedy materiał dotyczy partnera, szefa, dziecka, byłego małżonka, polityka budzącego nienawiść albo celebryty, którego ktoś uwielbia od piętnastu lat. Wtedy nie ogląda się wyłącznie obrazu. Ogląda się własne oczekiwania, uprzedzenia, lęki i nadzieje, tylko podane w formacie wideo, który przez dekady nauczył nas, że skoro coś zostało zarejestrowane, to musiało się wydarzyć.

Przez większą część współczesnej historii zdjęcie i nagranie miały szczególny status dowodu. Można było

kłamać w tekście, można było zmyślać podczas rozmowy, można było manipulować opisem wydarzenia, ale fotografia niosła ze sobą ciężar czegoś, co przynajmniej w jakimś momencie znalazło się przed aparatem. Film posiadał jeszcze większą władzę, ponieważ dodawał ruch, głos, czas i kontekst. "Przecież jest nagranie" kończyło wiele sporów skuteczniej niż dziesięć argumentów. Nie oznaczało to, że materiały wizualne były zawsze uczciwe. Kadrowano je, montowano, wycinano fragmenty i pozbawiano kontekstu od dawna. Mimo wszystko podstawowa umowa pozostawała stabilna: kamera mogła pokazać wydarzenie tendencyjnie, lecz zazwyczaj musiała mieć jakieś wydarzenie, które mogła pokazać. Deepfake rozrywa właśnie tę umowę, ponieważ pozwala stworzyć wiarygodny ślad zdarzenia, które nie musiało istnieć nawet przez sekundę.

Najbardziej kusząca reakcja brzmi więc: od tej chwili trzeba być bardziej sceptycznym. Tyle że sceptycyzm, który wydaje się ochroną przed fałszem, ma również swoją ciemniejszą wersję. Jeżeli można wygenerować dowolne nagranie, można również uznać dowolne prawdziwe nagranie za wygenerowane. Człowiek przytępiony na czymś kompromitującym dostaje prezent, o jakim wcześniejsze pokolenia mogły tylko marzyć. Nie musi już udowadniać, że materiał został zmanipulowany. Wystarczy, że zasieje wątpliwość. "To AI". "To deepfake". "Ktoś podłożył głos". "Každy potrafi dziś zrobić takie nagranie". W ten sposób technologia tworzenia fałszu zaczyna chronić również prawdę przed konsekwencjami, ponieważ prawdziwy materiał może zostać wrzucony do tego samego worka co tysiące syntetycznych podróbek. Największym sukcesem deepfake'u nie będzie więc dzień, w którym uwierzemy we wszystko. Znacznie wygodniejszy chaos powstanie wtedy, gdy przestaniemy być pewni czegokolwiek.

Ta niepewność zmienia bardzo codzienne sytuacje, nawet jeśli nie mają nic wspólnego z wyborami, celebrytami czy wielką geopolityką. Pracownik może otrzymać wiadomość głosową rzekomo od prezesa firmy z poleceniem wykonania pilnego przelewu. Partner może dostać nagranie sugerujące zdradę. Nastolatek może odkryć, że jego twarz została użyta w kompromitującym materiale krążącym po szkole. Nauczyciel może zobaczyć film, na którym podobno wypowiada słowa, których nigdy nie powiedział. Rodzina może zostać skonfrontowana z głosem bliskiej osoby proszącej o pomoc. W każdej z tych scen technologia jest ważna, ale jeszcze ważniejsza jest społeczna infrastruktura zaufania, na której opiera się reakcja. Głos matki ma znaczenie, ponieważ jest głosem matki. Twarz partnera boli, ponieważ należy do partnera. Polecenie szefa działa, ponieważ istnieje hierarchia. Deepfake nie tworzy tych więzi. On je wynajmuje na kilka minut i wykorzystuje przeciwko ludziom, którzy wcześniej budowali je latami.

Szczególnie brutalny jest moment, w którym ofiara próbuje się bronić i mówi: "To nie ja". Jeszcze niedawno takie zdanie można było porównać z materiałem i ocenić, czy zaprzeczenie ma sens. Teraz samo istnienie nagrania nie zamyka sprawy, ale brak pewności wcale nie oznacza, że społeczny wyrok zostanie odłożony. Internet nie ma zwyczaju czekać na ekspertyzę. Znajomi również nie zawsze mają cierpliwość do technicznej analizy, kiedy właśnie zobaczyli coś szokującego. Firma może reagować zanim pojawią się wyniki sprawdzenia, partner może wyprowadzić się zanim ktokolwiek ustali źródło filmu, a szkoła może już żyć materiałem, zanim osoba przedstawiona na nim zdąży wyjaśnić, że nigdy nie była w miejscu pokazanym w nagraniu. Technologia potrafi stworzyć fałsz w kilka minut, ale reputacja nie posiada przycisku przywracania ustawień

fabrycznych. Nawet po udowodnieniu manipulacji pozostaje drobne, brudne "a może jednak", które potrafi żyć znacznie dłużej niż sprostowanie.

W tej historii szczególnie niewygodne jest to, że człowiek uwielbia uważać własne zmysły za neutralnych świadków. Oczy widzą, uszy słyszą, mózg wyciąga wniosek i sprawa wydaje się zamknięta. Tymczasem percepcja nigdy nie działała jak bezstronny urzędnik zapisujący fakty w protokole. Mózg interpretuje, uzupełnia, porównuje z wcześniejszym doświadczeniem i bez przerwy decyduje, co uznać za istotne. Gdy widzi znajomą twarz i słyszy znajomy głos, nie rozpoczyna za każdym razem pełnego dochodzenia w sprawie ontologicznego statusu materiału. Rozpoznanie samo w sobie daje poczucie bezpieczeństwa poznawczego. Technologia, która potrafi naśladować właśnie te sygnały, trafia więc w system zaprojektowany nie do permanentnej podejrzliwości, lecz do szybkiego rozpoznawania świata. Człowiek przez tysiące lat nie musiał zakładać, że osoba stojąca przed nim może być matematycznie wygenerowanym zbiorem cech kogoś, kto znajduje się tysiąc kilometrów dalej.

Z tego powodu deepfake nie jest wyłącznie historią o tym, co potrafią komputery. Jest historią o tym, jak łatwo ludzkie przekonanie "widzę, więc wiem" może zostać oddzielone od rzeczywistości. Technologia po prostu brutalnie ujawnia, jak wiele pewności budowaliśmy na skrótach, których wcześniej nie trzeba było kwestionować przy każdym filmie. Gdy ktoś widział twarz znajomego człowieka, słyszał jego głos i obserwował ruch ust zsynchronizowany ze słowami, miał wystarczająco dużo powodów, żeby uznać materiał za prawdziwy. Teraz ten sam zestaw sygnałów może zostać skonstruowany. Nie oznacza to, że wszystko stało się fałszywe. Oznacza coś bardziej drażniącego: autentyczność

przestała być właściwością, którą można automatycznie wyczytać z samego wyglądu materiału.

Jeszcze bardziej interesujące zaczyna się wtedy, gdy fałszywe nagranie nie mówi niczego szczególnie szokującego. Człowiek spodziewa się przecież, że manipulacja będzie spektakularna, ponieważ spektakularność łatwo zauważyć. Tymczasem najskuteczniejszy materiał może być nudny. Dyrektor mówi spokojnie o przesunięciu terminu. Lekarz rzekomo przekazuje prostą informację. Partner wysyła krótką wiadomość głosową. Znana osoba wypowiada zdanie, które idealnie pasuje do jej wcześniejszych poglądów. Fałsz staje się wtedy częścią zwykłego przepływu informacji i nie aktywuje alarmu, ponieważ nie wygląda jak atak. Nie krzyczy, że chce manipulować. Zachowuje się jak kolejny element dnia, a właśnie takie elementy przepuszczamy przez uwagę z najmniejszym oporem. Człowiek sprawdza podejrzanе historie. Rzadziej sprawdza te, które brzmią dokładnie tak, jak powinny.

Powstaje więc paradoks człowieka uzbrojonego w coraz doskonalsze narzędzia komunikacji, który jednocześnie potrzebuje coraz więcej energii, żeby ustalić, czy komunikacja w ogóle pochodzi od osoby, której dotyczy. Kamera miała dokumentować. Mikrofon miał rejestrować. Wideorozmowa miała być bardziej bezpośrednia niż telefon. Nagranie miało kończyć dyskusję. Każde z tych narzędzi zwiększało poczucie obecności drugiego człowieka, aż technologia nauczyła się tę obecność syntetyzować. To niemal komiczny etap rozwoju cyfrowego świata: najpierw przez lata poprawialiśmy jakość obrazu, żeby wszystko wyglądało bardziej realistycznie, a potem musieliśmy zacząć zastanawiać się, czy właśnie dlatego realizm przestał

cokolwiek gwarantować. Kupiliśmy ekrany o coraz wyższej rozdzielczości, żeby zobaczyć więcej szczegółów, po czym pojawiły się narzędzia potrafiące wygenerować szczegóły, których nigdy nie było.

Cena tego przesunięcia nie jest wyłącznie technologiczna. Pojawia się koszt emocjonalny, kiedy człowiek zaczyna reagować na prawdopodobieństwo zamiast na pewność. Pojawia się koszt relacyjny, kiedy dowód nie kończy już konfliktu, tylko otwiera kolejny poziom podejrzeń. Pojawia się koszt tożsamościowy, kiedy własna twarz i własny głos przestają być czymś, nad czym ma się pełną kontrolę. Przez długi czas można było powiedzieć: "Nigdy bym czegoś takiego nie powiedział", a przynajmniej istniała granica między reputacją a fizycznym śladem zachowania. Teraz ktoś może stworzyć materiał, w którym człowiek mówi dokładnie to, czego nigdy by nie powiedział, używając jego sposobu artykulacji, mimiki i głosu. To szczególny rodzaj naruszenia, ponieważ nie kradnie jedynie danych. Kradnie wiarygodność ciała. Twarz staje się hasłem, które ktoś inny może próbować wpisać za ciebie.

Najbardziej niepokojące nie jest jednak to, że technologia potrafi tworzyć przekonujące fałszywe materiały. Bardziej niepokojące jest tempo, w jakim człowiek przyzwyczaja się do nowej rzeczywistości. Pierwsze zetknięcie z dobrym deepfakiem może wywołać szok. Dziesiąte staje się ciekawostką. Setne wpada między film z wakacji znajomego a reklamę butów. Psychika ma niezwykłą zdolność normalizowania rzeczy, które jeszcze chwilę wcześniej wydawały się nie do zaakceptowania. To mechanizm bardzo praktyczny, ponieważ bez niego każdy technologiczny przełom powodowałby trwały kryzys. Ma jednak swoją cenę: coś, co powinno zmienić sposób oceny informacji, może