Deepfake

Marcin Wilkowski

W 2019 roku w Internecie pojawiło się nagranie wideo wypowiedzi Marka Zuckerberga, założyciela Facebooka. Ostrzegał w nim przed sytuacją, kiedy jedna osoba ma całkowitą kontrolę nad danymi i sekretami miliardów ludzi. Problem w tym, że Zuckerberg nigdy niczego takiego nie powiedział.

Ujęcie trwa kilkanaście sekund. Zuckerberg siedzi przy biurku w beżowym tiszercie; wiadomo, że tak jak inni miliarderzy z Doliny Krzemowej unika noszenia garniturów i drogich ubrań. Mówiąc, gestykuluje – ruch dłoni i akcent podkreśla kluczowe słowa jego wypowiedzi: „skradzione dane", „kontrola", „sekrety". Nagranie nie było jednak świadectwem rewolucji w myśleniu o prywatności, którą Facebook zamienił w towar podlegający bezwzględnej komercjalizacji. Sfabrykowali je członkowie kolektywu artystycznego Big Dada Bill Posters i Daniel Howe, wykorzystując do tego metody uczenia maszynowego. Działający na ich podstawie program komputerowy użył oryginalnego nagrania telewizji CBSN z 2017 roku, które posłużyło za matrycę zmanipulowanego klipu. Twarz Zuckerberga została poddana automatycznej obróbce – algorytm dostosował ruchy ust i mimikę do słów fałszywej wypowiedzi.

Big Dada / Public Faces (2019) [dostęp 12.03.2021]. Dostępne w Bill Posters: http://bill-posters.ch/projects/big-dada/

↗ Ok. 3-minutowy film pt. *Big Dada/Public Faces* (2019) z udziałem Marcela Duchampa, Mariny Abramović, Marka Zuckerberga, Kim Kardashian, Morgana Freemana i Donalda Trumpa, wykonany techniką *deepfake*.

Akcja artystyczna kolektywu Big Dada podkreślała rosnące zagrożenie ze strony nagrań manipulowanych za pomocą algorytmów uczenia maszynowego.

Manipulacje w nagraniach filmowych nie są niczym nowym, pożądany efekt uzyskiwano choćby dzięki stosowaniu odpowiednich ujęć lub wycinaniu niektórych klatek już dawno temu, zanim film jako medium zaczął być nagrywany i edytowany cyfrowo. Rozwiązania informatyczne określane jako *deepfake* wznoszą jednak manipulacje wizualne na nowy poziom.

Naprawdę głębokie manipulacje

Deepfake to neologizm zbudowany z dwóch części. *Deep* pochodzi od terminu *deep learning*, głębokie uczenie maszynowe, które – gdyby ująć to w największym skrócie – umożliwia generowanie nowych treści na bazie zbiorów wykorzystywanych jako wzorzec, pozwalający algorytmowi się uczyć. Na przykład można użyć milionów fotografii ludzkich twarzy do wyuczenia algorytmu w taki sposób, aby był on w stanie generować zupełnie fikcyjne wizerunki. Podobnie w algorytmach generowania języka naturalnego bazą stają się miliony tekstów i to na ich podstawie algorytm uczy się tworzyć nowe treści – z powodzeniem rozwiązania te stosuje się choćby w edytorach tekstu podpowiadających piszącemu kolejne słowa albo w aplikacjach generujących krótkie teksty informacyjne o pogodzie, wypadkach lub wynikach sesji giełdowych.

Wymiarem skuteczności takich algorytmów jest jak największa naturalność stworzonych sztucznie treści, bez względu na to, czy chodzi o tekst informacji pogodowej, fotografię lub film pokazujący panoramę krajobrazu, który łączy oryginalne nagrania z zaawansowanym CGI (*computer-generated imagery*).

Fake akcentuje cele stosowania takich rozwiązań – w tym wypadku efektem wykorzystania głębokiego uczenia maszynowego ma być manipulacja i wprowadzenie w błąd odbiorców. O ile użycie takich narzędzi w kinie, w filmie fabularnym, nie jest niczym złym jako element kreacji świata przedstawionego, o tyle w odniesieniu do filmu dokumentalnego, wystąpienia politycznego czy relacji z miejsca zdarzenia może stać się nośnikiem propagandy lub wywoływać kryzysy polityczne bądź konflikty.

Manipulacje *deepfake* stosowano na przykład wobec kobiet, których dostępne w Internecie wizerunki wmontowywano do nagrań filmów pornograficznych – ofiarami były nie tylko gwiazdy i celebrytki, lecz także zwykłe użytkowniczki mediów społecznościowych. Proces generowania kilkuminutowego klipu tego rodzaju zajmował nawet ponad 40 godzin, ale można było go stworzyć na zwykłym domowym komputerze z użyciem darmowego oprogramowania, bez konieczności programowania czegokolwiek.

Deepfake jako metoda manipulacji jest więc groźny zarówno ze względu na naturalność uzyskiwanych efektów, jak i logikę komunikacji internetowej. Kiedy zalewa nas nieustannie potop informacji, tekstów i obrazów, brakuje czasu i energii na analizowanie ich autentyczności. Niekiedy w przypadku naturalizowanych manipulacji nie da się w kilka sekund rozpoznać fałszerstwa. Dodatkowo celem ataków z wykorzystaniem tej metody staje się na przykład medycyna, w tej dziedzinie bowiem analiza obrazu bywa już podstawą rozpoznawania niektórych chorób (choćby raka płuc). W 2019 roku wykonany w celach badawczych atak na jeden z ośrodków medycznych w Izraelu pozwolił na generowanie „deepfake'owych" skanów płuc w tomografii komputerowej – na szczęście nikt nie ucierpiał, bo fakt włamania został ujawniony przez samych jego sprawców. Podczas kampanii wyborczej w 2016 roku lekarz Hillary Clinton udostępnił online część jej dokumentacji medycznej w odpowiedzi na zarzuty, że stan zdrowia kandydatki nie pozwala na objęcie urzędu prezydenta USA. Głębokie manipulacje taką dokumentacją mogłyby więc mieć przełożenie także na procesy polityczne.

Deepfake w historii

Metody informatyczne pozwalające na generowanie tak głęboko zmanipulowanych treści opracowywane są w środowisku naukowym w ramach projektów badawczych, których celem jest zwiększenie możliwości pracy z obrazami cyfrowymi. Niekiedy mogą też mieć oryginalne naukowe zastosowania, również w dziedzinie badań przeszłości. W archeologii głębokie maszynowe analizowanie obrazów satelitarnych wspiera wyszukiwanie potencjalnych stanowisk lub rekonstruowanie uszkodzonych obiektów takich jak ceramika czy monety.

Shira Faigenbaum-Golovin i współautorzy w publikacji z 2016 roku wykazali użyteczność uczenia maszynowego w analizowaniu tekstów starożytnych zapisanych na fragmentach ceramiki około 600 roku p.n.e. – udało im się dowieść, że teksty wyszły spod ręki sześciu różnych autorów, dzięki automatycznemu porównywaniu liter z poszczególnych inskrypcji.

W naukach historycznych *deep learning* z powodzeniem stosuje się także w rozpoznawaniu pisma ręcznego w średniowiecznych manuskryptach lub do uzupełniania greckich inskrypcji, a metody CBIR (*content-based image retrieval*) umożliwiają przeszukiwanie olbrzymich zbiorów historycznych materiałów wizualnych – zamiast podawać słowa kluczowe jak w tradycyjnych wyszukiwarkach, użytkownik może wyszukiwać za pomocą obrazu.

Uczenie maszynowe umożliwia też wyodrębnianie wybranych fraz w dużych korpusach tekstów historycznych – na przykład automatyczne rozpoznawanie dat, miejsc, osób i czynności opisanych w zdigitalizowanych meldunkach wojskowych lub rozkazach dziennych – metoda ta określana jest jako NER (*named-entity recognition*). Włoscy badacze wykorzystali uczenie maszynowe do rozpoznawania wizerunków zniszczonych zabytków na historycznych nagraniach filmowych, wychodząc z założenia, że każda nowa informacja o nieistniejącym już obiekcie może być wartościowa i przydatna w budowaniu ich wirtualnej, trójwymiarowej rekonstrukcji.

Rozwiązania stojące za manipulacjami *deepfake* mogą służyć nowemu spojrzeniu na audiowizualne źródła historyczne. *Wjazd pociągu na stację w La Ciotat* to jeden z pierwszych niemych filmów zrealizowanych przez braci Lumière pod koniec XIX wieku. Ze względów historycznych nawet najlepsze zdigitalizowane wersje tego materiału mają niską jakość, jednak za pomocą uczenia maszynowego udało się to nagranie przeformatować do jakości 4K, dzięki czemu jego szczegóły są teraz bardziej widoczne. Była to właściwie operacja bardzo podobna do tych stosowanych w manipulacjach *deepfake* – nagranie zostało uzupełnione o nowe piksele na podstawie dostępnych bibliotek sztucznej inteligencji. Mamy więc tu do czynienia z reinterpretacją, a nie neutralną rekonstrukcją, co dla każdego badania historycznego powinno być fundamentalną różnicą. W podobny sposób powinniśmy traktować

czarno-białe fotografie historyczne koloryzowane automatycznie przez oprogramowanie działające na zasadzie uczenia maszynowego.

Andy Myers, *[Actual 4K Scan] The Arrival of a Train at La Ciotat Station - Lumière Brothers - 1896*, 2020 [dostęp 09.04.2021]. Dostępne w YouTube: https://www.youtube.com/watch?v=IFAj9fJQRZA

↑ Przyjazd pociągu na stację La Ciotat został nagrany przez braci Lumière w 1896 r. W lutym 2020 r. opublikowano odświeżoną i poddaną obróbce cyfrowej wersję, której jakość została podniesiona do poziomu 4K.

Odpowiedzialne podejście do sztucznej inteligencji wymaga też stawiania pytań o etykę podejmowanych działań. W przygotowywanym aktualnie filmie *Finding Jack* (reż. Anton Ernst i Tati Golykh) ma wystąpić zrekonstruowany wizerunek Jamesa Deana – nowe kwestie zostaną wgrane w zrekonstruowaną na bazie starych nagrań postać. Czy tego typu projekty nie są już przekroczeniem pewnej granicy?

W projekcie prowadzonym przez USC Shoah Foundation świadkowie Zagłady stają się wzorcem, na którym budowane są interaktywne hologramy: mechanizmy sztucznej inteligencji używane są nie tylko do ożywiania wirtualnej postaci, generowanej na podstawie obrazu z 52 kamer, lecz także do przygotowywania odpowiedzi na pytania zadawane przez rozmówców.

Opracowana specjalnie lista tysiąca pytań i odpowiedzi dotyczących biografii świadka i rzeczywistości Zagłady jest w tym projekcie bazą umożliwiającą dynamiczny, zbliżony do naturalnego dialog. Chociaż twórcy projektu podkreślają, że hologram bazuje na autentycznych wspomnieniach rzeczywistego świadka, który zgodził się wziąć udział w eksperymencie, to można się zastanawiać, czy nie pojawia się tu pole do pewnych nadużyć, szczególnie jeśli mechanizm generowania odpowiedzi proponowałby stwierdzenia bardzo mocno nadinterpretujące oryginalne wypowiedzi.

USC Shoah Foundation, Nimrod „Zigi" Ariav New Dimensions in Testimony, 2016 [dostęp 09.04.2021]. Dostępne w USC Shoah Foundation: https://sfi.usc.edu/gallery/nimrod-zigi-ariav-new-dimensions-testimony-nov-3-2016

← Świadek Zagłady Nimrod „Zigi" Ariav nagrywa swoje świadectwo w języku hebrajskim, Los Angeles, Kalifornia, kwiecień 2016 r.

Walka z zagrożeniem, jakim jest *deepfake*, polega na doskonaleniu metod jego rozpoznawania. Można analizować drobne błędy i niekonsekwencje w układzie scen, ruchu postaci czy mimice twarzy, można także, już na poziomie programistycznym, badać zmiany zachodzące między kolejnymi klatkami nagrania i odkrywać niespójności sugerujące manipulację.

Fałszerstwa da się wykryć także metodami statystycznymi, odwołując się na przykład do prawa Benforda. Prawo to określa prawdopodobieństwo występowania określonych pierwszych cyfr w szeregach danych liczbowych i stosowane jest już od dziesięcioleci, na przykład do wykrywania fałszerstw podatkowych. Okazuje się, że prawidłowości te wykazywać można również w danych z mediów społecznościowych czy strukturze plików komputerowych. Zaburzenie rozkładu Benforda może wskazywać na to, że plik graficzny czy nagranie wideo zostało zmanipulowane, nawet jeśli jego

warstwa wizualna wydaje się autentyczna. Być może wszystkie te metody i narzędzia staną się niedługo częścią warsztatu historycznego, wspierającego pracę ze źródłami cyfrowymi niemającymi pierwotnej postaci analogowej i dokumentującymi ważne wydarzenia historii najnowszej. Z pewnością jednak analiza techniczna tego typu zasobów powinna być poprzedzona podstawową krytyką źródła, sprawdzającą pochodzenie materiału, kontekst jego powstania czy autorstwo – takie podejście może być zresztą przydatne wobec każdego materiału rozpowszechnianego niezależnie w mediach społecznościowych.

Redakcja: Anna Kaniewska