AI Skills, Powered by Values

values

www.valuesai.eu

MODUŁ 03

Dezinformacja, deepfake'i oraz manipulacja danymi

Rozpoznawanie oraz zapobieganie oszustwom wykorzystującym sztuczną inteligencję

Co-funded by the European Union

Współfinansowane przez Unię Europejską. Wyrażone poglądy i opinie są wyłącznie poglądami i opiniami autora lub autorów i niekoniecznie odzwierciedlają stanowisko Unii Europejskiej ani Fundacji Rozwoju Systemu Edukacji. Ani Unia Europejska, ani instytucja przyznająca grant nie ponoszą za nie odpowiedzialności.

Zawartość

Co-funded by the European Union

Współfinansowane przez Unię Europejską. Wyrażone poglądy i opinie są wyłącznie poglądami i opiniami autora lub autorów i niekoniecznie odzwierciedlają stanowisko Unii Europejskiej ani Fundacji Rozwoju Systemu Edukacji. Ani Unia Europejska, ani instytucja przyznająca grant nie ponoszą za nie odpowiedzialności.

01

Dezinformacja

(20 minut)

Dezinformacja

Dezinformacja wspierana przez sztuczną inteligencję odnosi się do celowo fałszywych lub wprowadzających w błąd treści, które są tworzone, edytowane lub rozpowszechniane za pomocą narzędzi sztucznej inteligencji, takich jak modele generatywne, deepfake'i czy boty, w celu oszukania opinii publicznej, wpływania na postrzeganie oraz wyrządzania szkody procesom demokratycznym, instytucjom lub jednostkom.

AI Skills,
Powered
by Values
values

Przykład dezinformacji.

Jednym z najsłynniejszych i najlepiej udokumentowanych przykładów dezinformacji wspieranej przez sztuczną inteligencję jest **fałszywy** obraz aresztowania Donalda Trumpa, który w 2023 roku zyskał popularność w mediach społecznościowych oraz serwisach informacyjnych w Europie i Stanach Zjednoczonych.

Na zdjęciu widać, jak policja prowadzi Trumpa, mimo że **do aresztowania nigdy nie doszło** – obraz został **w pełni wygenerowany przez narzędzie do tworzenia obrazów oparte na sztucznej inteligencji**, bazując na rzeczywistych archiwalnych nagraniach i zdjęciach.

Wiele osób, które udostępniły te zdjęcia, zauważyło, że są one fałszywe - nie udało im się oszukać wielu ludzi, jednak kilka osób rzeczywiście dało się zwieść.

Dyskusja

Będziemy świadomi, że nasz program dezinformacyjny został zakończony, gdy wszystko, w co wierzy amerykańska opinia publiczna, okaże się fałszem.

William J. Casey, dyrektor Centralnej Agencji Wywiadowczej

1. Zastanów się nad potencjalnymi konsekwencjami ekonomicznymi i społecznymi, jakie może wywołać taka wiadomość.
2. Czy szkody wynikające z takiej działalności są całkowicie odwracalne?
3. Medium niosącym zdecydowanie największe ryzyko dezinformacji jest internet. Zastanów się, dlaczego tak się dzieje.
4. Przeprowadź głosowanie w grupie, aby ustalić, czy bardziej ufasz wiadomościom telewizyjnym, czy informacjom z internetu. W jaki sposób wynik Twojego głosowania koreluje z ryzykiem dezinformacji?

Błędna informacja, dezinformacja i manipulacja to nie to samo.

Błędna informacja

Fałszywe lub nieprecyzyjne informacje nie zawsze są rozpowszechniane intencjonalnie.

Dezinformacja

Fałszywe informacje celowo generowane i rozpowszechniane

Manipulacja

Zmiana treści, kontekstu lub danych w celu oddziaływania na odbiorcę

02

Deepfake'i

(20 minut)

Deepfake'i: Dlaczego są trudne do wykrycia

- **Deepfake to materiał medialny, który został stworzony lub zmodyfikowany przy użyciu sztucznej inteligencji, aby sprawiać wrażenie, że dana osoba powiedziała lub zrobiła coś, czego w rzeczywistości nie uczyniła.** Zaawansowane algorytmy potrafią łączyć autentyczne nagrania z obrazami i dźwiękiem generowanymi komputerowo w sposób tak płynny, że często trudno to rozpoznać podczas oglądania lub słuchania.

- **Technologia ta rozwija się w zawrotnym tempie**. Nie pozostawia już widocznych błędów ani „znaków rozpoznawczych”, takich jak niedopasowane usta czy rozmyte tło. W związku z tym nawet doświadczeni dziennikarze i eksperci w dziedzinie informatyki śledczej mogą mieć trudności z identyfikacją starannie wykonanej podróbki.

Deepfake'i: Dlaczego są trudne do wykrycia

- **Deepfake'i mogą być wykorzystywane w celach humorystycznych lub artystycznych, jednak coraz częściej służą do wprowadzania w błąd lub wyrządzania szkód.** Mogą zaszkodzić reputacji, rozpowszechniać dezinformację lub być używane w atakach politycznych czy personalnych. Ich realizm sprawia, że musimy rozwijać w sobie ostrożny i dociekliwy sposób myślenia, gdy napotykamy zaskakujące treści w internecie.

Czy chcesz zweryfikować swoje umiejętności detektywistyczne w obszarze cyfrowym?

Strona internetowa Detect Fakes MIT Media Lab pozwala na oglądanie autentycznych filmów oraz filmów stworzonych przez sztuczną inteligencję, a następnie na podejmowanie decyzji, które z nich są które.

Naukowcy opracowali to narzędzie, aby wspierać ludzi w nauce poprzez praktykę – nie istnieje jeden uniwersalny sposób na identyfikację deepfake'a, jednak zwracając uwagę na subtelne detale twarzy i oświetlenia, można zacząć dostrzegać, kiedy coś zostało zmanipulowane.

Twój zestaw narzędzi do autorefleksji

Podczas tej sesji wystąpią chwile, w których poprosimy Cię o zatrzymanie się i refleksję nad tym, co właśnie zobaczyłeś lub usłyszałeś. Zachęcamy do sporządzania notatek w notesie, telefonie lub aplikacji do notowania.

ieje jedna, właściwa odpowiedź; te notatki są czone do Twojego użytku i mają na celu wspieranie ykształceniu nawyku kwestionowania tego, co ujesz w Internecie. Celem jest rozwinięcie ności cyfrowego detektywa oraz nauczenie się, jak radzić sobie z niepewnością.

01

Jakie przesłanki mogą sugerować, że film lub zdjęcie jest fałszywe?

02

Jakie myśli lub pytania nasuwają się w Twojej głowie, szczególnie gdy coś wydaje się nie w porządku?

03

Dlaczego ktoś mógłby stworzyć i udostępnić fałszywy produkt?

Co o tym sądzisz?

Zrozumienie, czym są deepfake’i oraz dlaczego ich wykrycie jest trudne, umożliwi ci skorzystanie z trzypunktowej listy kontrolnej, którą znajdziesz w dalszej części tego modułu.

01

Czy kiedykolwiek natknąłeś się na deepfake lub podejrzewałeś, że film został zmanipulowany? Jakie były Twoje odczucia w tej sytuacji?

02

Wiedząc, jak autentyczne mogą wydawać się takie podróbki, jakie dodatkowe kroki powinieneś podjąć, zanim uwierzysz w szokujący filmik lub go udostępnisz?

03

W jaki sposób można pozytywnie wykorzystać technologię deepfake oraz jak możemy zapobiegać jej niewłaściwemu zastosowaniu?

Pomyśl: poruszanie się w niepewności

Poniższe ćwiczenia mają na celu rozwijanie umiejętności krytycznego myślenia oraz ułatwienie poruszania się w cyfrowym świecie, w którym fakty i fikcja często się przenikają.

01

Jak szybko dzielisz się interesującym lub zabawnym obrazem lub klipem AI? Jak podejmujesz decyzję o jego udostępnieniu?

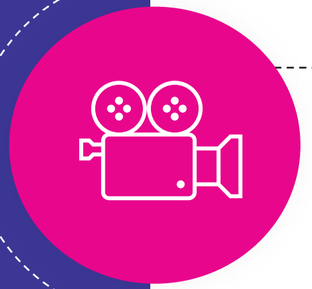
02

Jaka jest różnica między zabawnym obrazem lub klipem AI a takim, który szkodzi ludziom?

03

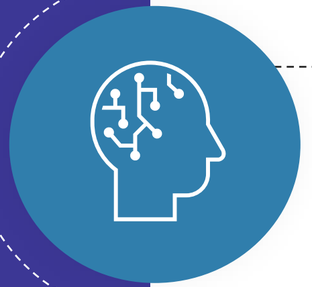
Kto może odnieść korzyści z rozprzestrzeniania dezinformacji na temat sztucznej inteligencji?

Rodzaje fałszywych treści generowanych przez sztuczną inteligencję – definicje oraz przykłady



01 **Fałszywe filmy wideo** – nagrania, na których osoba wydaje się mówić lub wykonywać czynności, których w rzeczywistości nigdy nie powiedziała ani nie zrobiła, stworzone lub zmodyfikowane przy użyciu sztucznej inteligencji (np. deepfake).

02 **Fałszywe komentarze i recenzje** – opinie, oceny lub komentarze tworzone przez sztuczną inteligencję w Internecie, mające na celu sztuczne kreowanie pozytywnego wizerunku produktu, firmy lub polityka.



03 **Konta syntetyczne i boty** – fikcyjne profile użytkowników lub konta botów zarządzane przez sztuczną inteligencję, które naśladują rzeczywistych użytkowników w celu zwiększenia zasięgu nieprawdziwych treści.

04 **Kampanie wpływu generowane automatycznie** – kampanie informacyjne, w których treść (teksty, grafiki, wideo, posty w mediach społecznościowych) jest w całości lub częściowo tworzona przez sztuczną inteligencję i zautomatyzowana w celu masowego rozpowszechniania.

Rodzaje fałszywych treści generowanych przez sztuczną inteligencję – definicje oraz przykłady

05 fałszywe wiadomości – artykuły lub nagłówki w mediach, które są generowane lub modyfikowane przez sztuczną inteligencję, aby przypominały prawdziwe wiadomości, lecz w rzeczywistości zawierają nieprawdziwe lub zniekształcone informacje

06 fałszywe cytaty – stwierdzenia błędnie przypisywane celebrytom, politykom lub naukowcom, generowane przez sztuczną inteligencję w postaci „cytatów”

07 fałszywe obrazy – grafiki lub „zdjęcia” wydarzeń, które nigdy nie miały miejsca (np. aresztowania polityczne, katastrofy, wydarzenia społeczne), wytwarzane przez generatory obrazów AI

08 fałszywe nagrania głosowe – syntetyczne głosy osób publicznych stosowane do generowania nieautentycznych klipów audio, rozmów, oświadczeń lub przemówień

03

Manipulacja danymi oraz kontekstem

(15 minut)

Manipulacja danymi treści zazwyczaj odnosi się do:

- **Selektywnej prezentacji danych** – ukazywanie jedynie tej części zbioru danych, która wspiera dane twierdzenie, przy jednoczesnym pomijaniu pozostałych elementów.
- **Zmiany skali wykresu** – zastosowanie skróconej osi lub nierównych interwałów, co sprawia, że niewielkie różnice wydają się znacznie bardziej wyraźne.
- **Pominięcia źródła** – prezentacja liczby, wykresu lub cytatu bez podania źródła, metodologii lub daty.
- **Fałszywych statystyk** – prezentowanie wymyślonych, nieprawidłowych lub matematycznie niepoprawnych danych liczbowych.

Manipulacja danymi treści zazwyczaj odnosi się do:

- **Wyrywania wypowiedzi z kontekstu** – prezentowania jedynie krótkiego fragmentu w sposób, który zmienia pierwotne znaczenie.
- **Łączenia autentycznych danych z błędną interpretacją** – stosowania właściwych danych, lecz formułowania na ich podstawie mylących lub manipulacyjnych wniosków.
- **Selektywnego wyboru przedziału czasowego** – wybierania daty początkowej i końcowej w sposób, który sztucznie sugeruje wzrost, spadek lub kryzys danych.
- **Manipulacyjnego wyboru kategorii lub próby** – grupowania danych lub wyboru próby w sposób, który sprawia, że wynik wydaje się bardziej przekonujący, niż jest w rzeczywistości.

Przykłady praktyczne

Ukrywanie tendencji

Załóżmy, że chcemy przedstawić wartości sprzedaży czterech produktów oferowanych przez naszą firmę. Niestety, produkt B (kolor czerwony), który stanowi fundament naszej oferty, nie sprzedaje się tak dobrze jak wcześniej. Nie ma problemu – **możemy to z łatwością ukryć.**

Używając wykresu słupkowego skumulowanego, możemy sprawić, że spadek sprzedaży pozostanie niezauważony lub przynajmniej będzie wydawał się mniej drastyczny niż w rzeczywistości. Wizualizacja tych samych danych na wykresie liniowym pozwala nam wyraźnie dostrzec rzeczywisty obraz sytuacji – trend spadkowy jest wyraźnie widoczny. **Trudno uwierzyć, że oba wykresy przedstawiają te same dane, prawda?**

Zobacz kolejny slajd

Ukrywanie tendencji

ICAN. (b.d.). Krótki przewodnik po manipulacji w wizualizacjach danych. Pobrano 17 maja 2026 r. z https://www.ican.pl/b/krotki-przewodnik-po-manipulacji-wizualizacjami-danych/16IwGIn0h

Przykłady praktyczne

Zmiana skali

Na pierwszym wykresie skala rozpoczyna się od zera. Możemy dostrzec, że różnica między słupkami 2 a 7 nie jest istotna.

Na drugim wykresie ponownie przedstawione są te same dane. W tym przypadku skala rozpoczyna się od 100. Różnice między poszczególnymi słupkami są tutaj istotne.

Zobacz kolejny slajd

Zmiana skali

Przykłady praktyczne

Fałszywa perspektywa

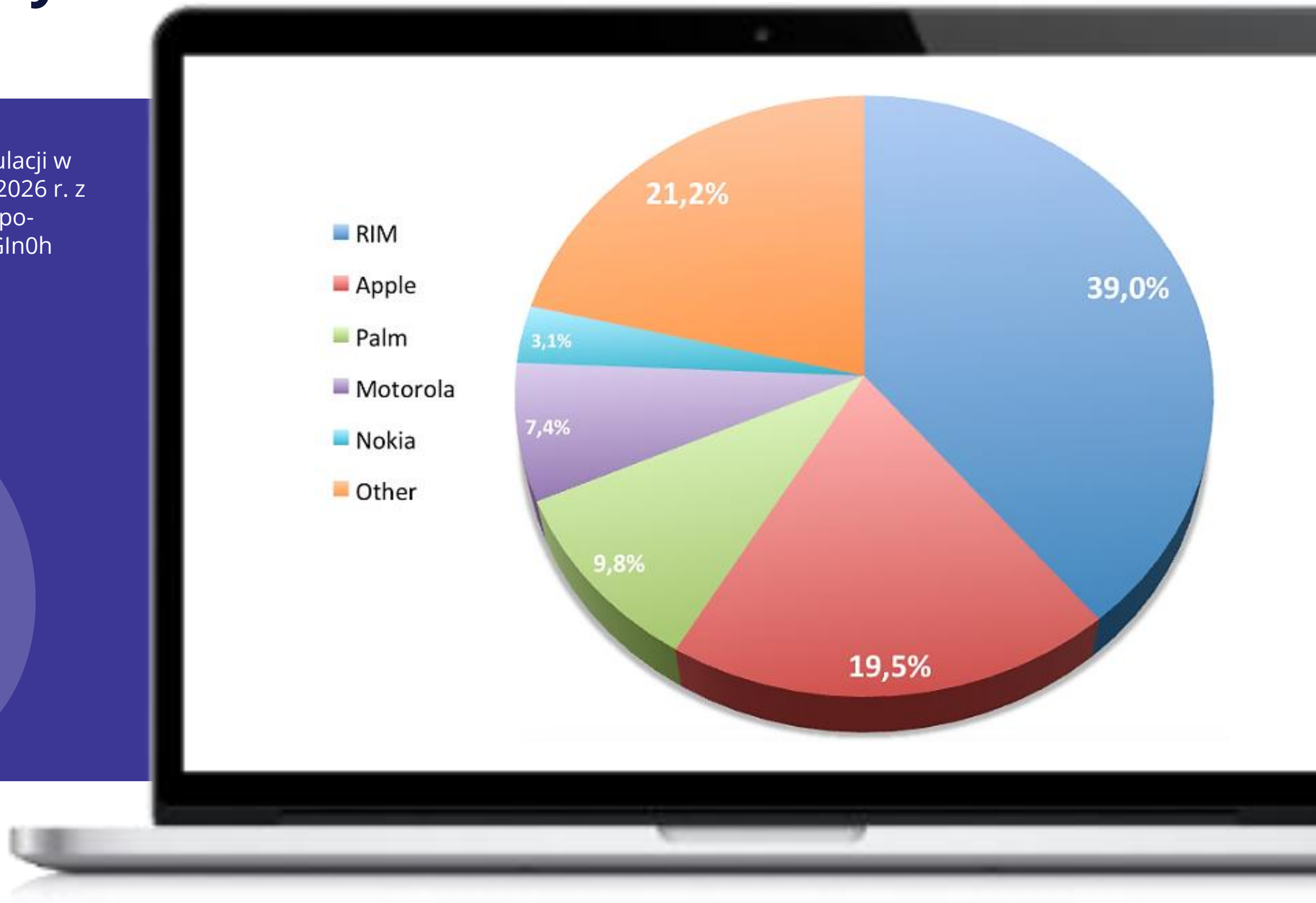
W 2008 roku Steve Jobs zaprezentował udział Apple w rynku smartfonów. Na powyższym wykresie (który jest wierną reprodukcją tego przedstawionego przez Steve'a Jobsa) wyraźnie widać, że segment reprezentujący Apple jest większy niż ten reprezentujący „innych”. Nie wydaje się on również znacząco mniejszy od udziału RIM (BlackBerry). Jednak analizując wartości procentowe, Apple posiadało wówczas jedynie połowę udziału RIM w rynku.

Na pierwszy rzut oka nie widać, że to wykres trójwymiarowy. Został on sprytnie obrócony, aby sprawiał wrażenie bardziej pionowego. Efekt ten jest dodatkowo wzmocniony przez zmianę rozmiaru czcionki: mniejsza czcionka jest stosowana dla akcji o mniejszym znaczeniu, natomiast ta sama wielkość dla RIM, Apple i „innych”.

Zobacz kolejny slajd

Fałszywa perspektywa

ICAN. (b.d.). Krótki przewodnik po manipulacji w wizualizacjach danych. Pobrano 17 maja 2026 r. z https://www.ican.pl/b/krotki-przewodnik-po-manipulacji-wizualizacjami-danych/16IwGIn0h



Przykłady praktyczne

Zręczne grupowanie

Wykresy ilustrują system podatkowy w Stanach Zjednoczonych. Na wykresie po lewej stronie możemy dostrzec dochody (w wartościach bezwzględnych) różnych grup w zależności od poziomu zarobków.

Subtelna manipulacja polega na klasyfikowaniu wyników w różne kategorie, aby wykres przyjął pożądany kształt. W ten sposób wykres przypominający rozkład normalny (po lewej) można przekształcić w wykres w kształcie kija hokejowego (po prawej).

Przykłady praktyczne

Zręczne grupowanie

Patrząc na wykres po lewej stronie, dochodzimy do wniosku, że osoby z klasy średniej dysponują największymi zasobami finansowymi. Na wykresie po prawej stronie kilka kategorii zostało zintegrowanych w jedną większą grupę.

W rezultacie możemy stwierdzić, że klasa wyższa (od 200 tysięcy dolarów wzwyż) dysponuje największymi zasobami. Te same dane mogą prowadzić do odmiennych wniosków, a w konsekwencji do różnych decyzji obywatelskich, w tym wyborczych.

Zobacz kolejny slajd

Zręczne grupowanie

ICAN. (b.d.). Krótki przewodnik po manipulacji w wizualizacjach danych. Pobrano 17 maja 2026 r. z https://www.ican.pl/b/krotki-przewodnik-po-manipulacji-wizualizacjami-danych/16IwGIn0h

04

Twój zestaw narzędzi śledczego

(20 minut)


Sztuczna inteligencja jest zdolna do generowania przekonujących fałszywych mediów, dlatego do oceny autentyczności konieczne jest zastosowanie ustrukturyzowanego podejścia.
AI Skills, Powered by Values
values

Lista kontrolna, która pomoże Ci ukierunkować myślenie

Te wskazówki nie są niezawodne; raczej pomagają ci zwolnić tempo i dokładniej przyjrzeć się sytuacji, zanim podzielisz się tym, co dostrzegasz, lub uwierzysz w to, co widzisz.

1. **Sprawdź szczegóły**: **Zwróć uwagę na detale**: czy dana osoba ma właściwą liczbę palców? Czy tła lub odbicia wydają się zniekształcone? Czy oświetlenie wygląda naturalnie? Deepfake'i często wprowadzają w błąd w tych aspektach.
2. **Sprawdź źródło**: **Zadaj sobie pytanie:** kto to opublikował lub udostępnił? Czy pochodzi z wiarygodnego serwisu informacyjnego, zaufanej osoby, czy może jest to nieznane źródło? Spróbuj zweryfikować, czy rzetelne źródła przedstawiają tę samą historię.
3. **Sprawdź swoją intuicję**: Zauważ, jakie emocje wywołuje w Tobie dana treść. Czy jej celem jest szokowanie, wzbudzanie gniewu lub prowokowanie? Jeśli coś wydaje się zbyt sensacyjne, aby mogło być prawdziwe, zatrzymaj się i zweryfikuj, zanim zareagujesz.

Skorzystaj z tej listy kontrolnej za każdym razem, gdy napotkasz zaskakujące treści. Pamiętaj, że nawet eksperci czasami mają trudności z odróżnieniem prawdy od fałszu, a zawsze lepiej jest zweryfikować informacje dwukrotnie niż szerzyć dezinformację.

Sprawdź szczegóły.

- **Filmy i obrazy generowane przez sztuczną inteligencję** mogą wydawać się przekonujące, jednak drobne nieścisłości często je zdradzają. Pierwsza wskazówka koncentruje się na dostrzeganiu tego, co wydaje się „nie tak”, zanim jeszcze zrozumiesz, dlaczego.
- **Twoja pierwsza umiejętność detektywistyczna**: zwolnij tempo i przyjrzyj się uważniej. Deepfake'i często pomijają detale, które w rzeczywistości nigdy nie umykają uwadze.
- **Sztuczna inteligencja potrafi naśladować twarze, głosy i gesty, jednak wciąż napotyka trudności z drobnymi szczegółami.** Te błędy mogą być subtelne: dziwne odbicie w okularach, dodatkowy palec, nieodpowiadające cienie czy kolczyki zmieniające strony między zdjęciami. Umiejętność ich dostrzegania stanowi pierwszy krok do rozwinięcia cyfrowej intuicji.

Sprawdź źródło.

- Zidentyfikuj oryginalnego autora pliku lub witrynę, na której jest on umieszczony.
- Zweryfikuj, czy rzetelne serwisy informacyjne lub oficjalne źródła przedstawiają tę samą narrację.
- Skorzystaj z odwrotnego wyszukiwania obrazów, aby zweryfikować, czy zdjęcie występuje w innych kontekstach.
- Pamiętaj, że pierwsze źródło, które rozważysz, rzadko bywa najbardziej wiarygodne.

Sprawdź źródło.

Nawet najbardziej realistyczny obraz lub film może wprowadzać w błąd, jeśli pochodzi z niewiarygodnego źródła. Druga wskazówka koncentruje się na źródle treści oraz metodach jej weryfikacji.

Łatwo zapomnieć, że każdy post, obraz czy film ma swojego twórcę: osobę, która zdecydowała się go udostępnić. Deepfake lub wprowadzający w błąd obraz może być przesłany przez fałszywe konto lub ponownie udostępniony przez kogoś, kto nigdy go nie zweryfikował. Dobrzy detektywi podążają za tropem.

Zaufaj przeczuciu

Nie musisz od razu wiedzieć, czy coś jest prawdziwe, ale zawsze możesz rozważyć, jakie uczucia to w tobie wywołuje.

- Czasami najlepszym narzędziem, które posiadasz, jest Twój własny instynkt. Ta wskazówka wspiera uczniów w rozpoznawaniu, kiedy ich emocje są manipulowane: kluczowa umiejętność dla cyfrowej odporności.
- Jeśli coś wydaje się zbyt doskonałe, zbyt szokujące lub zbyt emocjonalne, może to być zamierzone, aby wywołać u ciebie reakcję, zanim zdążysz się zastanowić.
- Treści generowane przez sztuczną inteligencję często oddziałują na nasze emocje. Mogą nas rozwścieczyć, zainspirować lub przestraszyć: emocje skłaniają ludzi do klikania, udostępniania i wierzenia.

Zaufaj przeczuciu

Zanim coś udostępnisz dalej, pomyśl o spowolnieniu reakcji emocjonalnej.

Kiedy coś wywołuje intensywną reakcję, zatrzymaj się. Krytyczne myślenie nie opiera się na cynizmie, lecz na ciekawości. Nie musisz od razu wiedzieć, czy coś jest prawdziwe, ale zawsze możesz zastanowić się, jakie emocje w tobie budzi.

Rozważasz publikację treści w Internecie?

Sprawdź te pięć kroków, aby w pierwszej kolejności zadbać o bezpieczeństwo swoje i innych w sieci!

Dlaczego warto korzystać z krytycznego myślenia w internecie?

Badania wskazują, że...

Nawet krótka ekspozycja na nieprawdziwe treści może kształtować przekonania.
(Pennycook i Rand, 2019)

Treści emocjonalne rozprzestrzeniają się szybciej niż poprawki faktów.

(Vosoughi i wsp., 2018)

Samo obalanie czegoś nie eliminuje wpływu emocjonalnego.
(Nyhan i Reifler, 2010)

Narzędzia oraz metody weryfikacji faktów

1. **Odwrotne wyszukiwanie obrazów –** analiza obrazu w odwrotnej kolejności w celu zidentyfikowania jego pierwotnego źródła, wcześniejszego zastosowania oraz weryfikacji, czy nie został wyrwany z kontekstu.
2. **Sprawdzanie metadanych –** analiza ukrytych danych pliku, takich jak data utworzenia, urządzenie, lokalizacja lub autor, o ile są dostępne.
3. **Porównywanie z oficjalnymi kanałami** – weryfikacja treści w odniesieniu do oświadczeń instytucji, organizacji, zweryfikowanych kont lub stron internetowych źródłowych.
4. **Portale weryfikacji faktów** – korzystanie z witryn internetowych, które sprawdzają dokładność twierdzeń, obrazów i filmów.

Narzędzia oraz metody weryfikacji faktów

5. **Weryfikacja źródeł pierwotnych** – śledzenie informacji do ich oryginalnego źródła, zamiast opierania się na powtórzeniach, cytatach i podsumowaniach.
6. **Weryfikacja daty i miejsca publikacji** – analiza, kiedy i gdzie materiał został opublikowany oraz czy jest zgodny z opisanym wydarzeniem.
7. **Analiza kontekstu** – badanie całościowego kontekstu komunikacyjnego w celu identyfikacji cytatów, obrazów lub filmów, które zostały oderwane od swojego pierwotnego znaczenia.
8. **Geolokalizacja wizualna** – porównywanie elementów na zdjęciu lub filmie z mapami, budynkami, znakami, krajobrazami lub warunkami atmosferycznymi w celu ustalenia lokalizacji, w której wykonano zdjęcie lub nagranie.

05

Odpowiedzialność twórców treści oraz skutki dezinformacji

(15 minut)

Odpowiedzialność autora treści

1. **Nie publikuj niezweryfikowanych informacji** – zanim cokolwiek udostępnisz, upewnij się, że dane są prawdziwe i potwierdzone przez kilka wiarygodnych źródeł.
2. **Nie udostępniaj treści, których autentyczność jest wątpliwa** – jeśli nie masz pewności co do prawdziwości materiału, nie rozpowszechniaj go.
3. **Oznaczaj treści syntetyczne** – jednoznacznie informuj, kiedy obraz, dźwięk, tekst lub wideo zostały stworzone lub zmodyfikowane przez sztuczną inteligencję.
4. **Poprawiaj błędy** – jeśli dostrzeżesz błąd w opublikowanej treści, skoryguj ją jak najszybciej i w sposób wyraźny.
5. **Usuń lub popraw wprowadzające w błąd treści** – jeśli treść może mylnie sugerować fakty, które nigdy nie miały miejsca, należy ją usunąć lub wyjaśnić.

Odpowiedzialność autora treści

6. **Nie używaj czyjegoś wizerunku ani głosu bez jego zgody –** wykorzystywanie twarzy, głosu lub stylu mówienia innej osoby bez jej zgody może naruszać prywatność oraz prawa osobiste.
7. **Podaj źródła oraz kontekst** – wyjaśnij, skąd pochodzi materiał, kiedy został stworzony i dlaczego został opublikowany.
8. **Nie manipuluj świadomie odbiorcami –** twórcy powinni unikać działań, które celowo wprowadzają odbiorców w błąd w celu uzyskania korzyści, zwiększenia zasięgu lub wywołania emocji.

Wpływ wprowadzającej w błąd sztucznej inteligencji.

Fałszywe lub wprowadzające w błąd treści generowane przez sztuczną inteligencję nie dotyczą wyłącznie jednej osoby.

Gdy fałszywy obraz, film lub opowieść zostanie udostępniona, rozprzestrzenia się błyskawicznie na wielu platformach, docierając do osób, które ufają nadawcy, i może wyrządzić szkody wykraczające poza pierwotny wpis.

„Badacze określają to jako efekt domina: jedna nieprawdziwa informacja może znacząco wpłynąć na zaufanie, zachowanie i decyzje w szerokim zakresie.”

(Wardle i Derakhshan, 2017).

Rodzaje strat

Fałszywe informacje generowane przez sztuczną inteligencję mogą powodować:

- **Szkody osobiste**: szkoda dotycząca reputacji lub zdrowia psychicznego danej osoby.
- **Szkody społeczne**: lęk, konflikt lub dyskryminacja
- **Szkody demokratyczne**: wprowadzanie wyborców w błąd lub podważanie zaufania do instytucji.
- **Szkody dotyczące bezpieczeństwa**: oszustwa, wyłudzenia lub nieprawdziwe informacje medyczne

Rodzaje strat

Nawet jeśli dalsze udostępnianie wydaje się niewinne, jego konsekwencje mogą być poważne.
Obejrzyj „Bad science: AI used to target kids with disinformation on YouTube - BBC World Service”.

Kliknij, aby obejrzeć

06

Ćwiczenia oraz przykłady

(25 minut)

Ćwiczenie: Odgrywanie ról – dylemat zaawansowanej podróbki

Scenariusz 01

W mediach społecznościowych w ramach żartu opublikowano deepfake przedstawiający kolegę z klasy — co powinni uczynić przyjaciele?

Scenariusz 02

Popularny influencer internetowy publikuje film ze swoim partnerem/partnerką – później okazuje się, że partner/partnerka nie istnieje i został stworzony wyłącznie przy użyciu sztucznej inteligencji, aby zwiększyć oglądalność. Fani czują się oszukani.

Ćwiczenie: Odgrywanie ról – dylemat zaawansowanej podróbki

Scenariusz 03

Po bolesnym rozstaniu ktoś publikuje w internecie nagranie, na którym jego były partner wypowiada okrutne słowa lub przyznaje się do zdrady. Film szybko się rozprzestrzenia, lecz w rzeczywistości jest to deepfake stworzony przez sztuczną inteligencję.

Scenariusz 04

Odkrywasz, że ktoś wykorzystał sztuczną inteligencję do naśladowania Twojego głosu w filmie-żarcie —
Jak sobie z tym radzisz?

„Papież w puchowej kurtce”

Fałszywe zdjęcia papieża Franciszka w puchowej kurtce zyskują popularność w sieci, ukazując zarówno moc, jak i zagrożenie związane ze sztuczną inteligencją.

„Papież w puchowej kurtce”

- **Co się wydarzyło**: Wizerunek papieża Franciszka w białej puchowej kurtce, stworzony przez sztuczną inteligencję, zyskał status wiralowego; wiele osób uważało go za autentyczny.
- **Co wprowadza w błąd**: Fotorealizm, wskazówki dotyczące trendów mody.
- **Intencja**: Humorystyczna
- **Szkoda**: Podważa zaufanie, normalizuje brak weryfikacji źródeł.

Jakie subtelne wskazówki możesz zauważyć na tym obrazku, zanim go udostępnisz?

Filtr TikTok „Odważny Urok”

Filtr upiększający Bold Glamour z TikToka budzi niepokój wśród ekspertów ds. sztucznej inteligencji w kontekście rzeczywistości.

The Washington Post

Filtr TikTok „Odważny Urok”

- **Co się wydarzyło:** Hiperrealistyczny filtr upiększający wykorzystuje algorytmy uczenia maszynowego do modyfikacji kształtu twarzy w czasie rzeczywistym.
- **Co wprowadza w błąd**: Płynna edycja twarzy za pomocą sztucznej inteligencji wydaje się „naturalna”.
- **Intencja**: Rozrywka/angażowanie.
- **Szkody**: nierealistyczne normy, presja związana z percepcją własnego ciała.

Jak ten filtr może wpływać na postrzeganie siebie i innych przez ludzi?

„Dyrektor finansowy w technologii deepfake na Zoomie”

Pracownik firmy w Hongkongu zapłacił 20 mln funtów w ramach oszustwa związanego z wykorzystaniem technologii deepfake w rozmowach wideo.

Hongkong | Guardian

„Dyrektor finansowy w technologii deepfake na Zoomie”

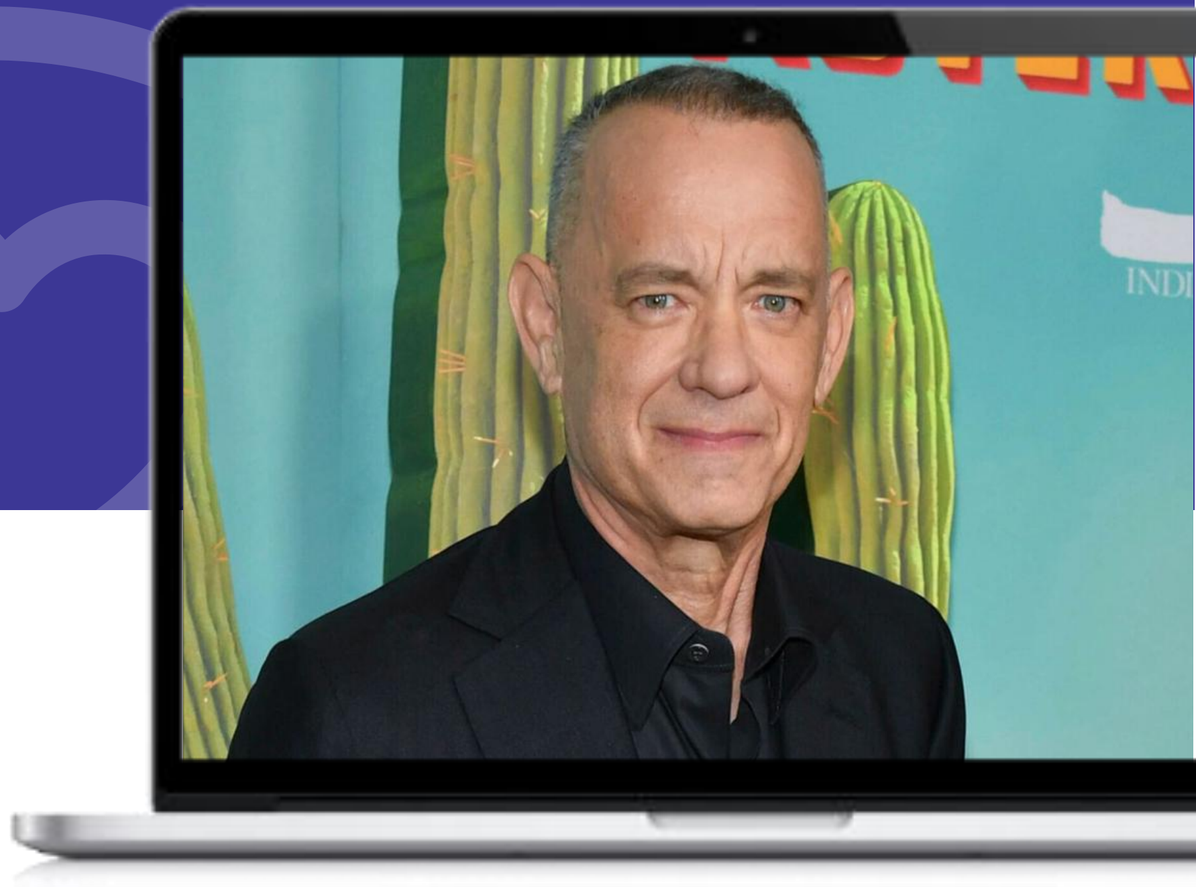
- **Co się wydarzyło**: Pracownik działu finansowego w Hongkongu padł ofiarą oszustwa, uczestnicząc w wieloosobowej, fałszywej rozmowie wideo z rzekomymi „współpracownikami”, w wyniku czego przekazał łącznie około 25 mln dolarów.
- **Co wprowadza w błąd**: Przekonujące klony nagrań wideo oraz głosów znanych pracowników.
- **Intencja**: Oszustwo.
- **Szkoda**: znaczne straty finansowe, zagrożenie dla reputacji oraz ryzyko prawne.

Podaj dwie rodzaje kontroli, które będą niezbędne przed każdym transferem.

„Ta reklama to nie ja”

Tom Hanks ostrzega fanów, aby nie dali się oszukać reklamą deepfake, która wykorzystuje jego wizerunek.

Wiadomości z obszaru rozrywki i sztuki | Sky News



„Ta reklama to nie ja”

- **Co się stało**: Film stworzony przez sztuczną inteligencję wykorzystał wizerunek Toma Hanksa do promocji produktu; Tom Hanks ostrzegł swoich fanów.
- **Co wprowadza w błąd**: Zaufana osoba oraz głos, format reklamy.
- **Intencja**: Oszustwo handlowe.
- **Szkoda**: Oszustwa konsumenckie, niewłaściwe wykorzystanie praw do wizerunku.

Jak weryfikujesz rekomendację renomowanej osoby, zanim jej uwierzysz?

„Nie głosuj dzisiaj.”

Konsultant polityczny odpowiedzialny za automatyczne połączenia telefoniczne Bidena generowane przez sztuczną inteligencję stoi w obliczu grzywny w wysokości 6 milionów dolarów oraz zarzutów karnych.

„Nie głosuj dzisiaj”

- **Co się wydarzyło**: Głos sztucznej inteligencji, naśladujący głos prezydenta Bidena, wezwał wyborców do opuszczenia prawyborów; w konsekwencji nałożono grzywny oraz postawiono zarzuty.
- **Co wprowadza w błąd**: Znajomy głos polityczny, zaprezentowany przed głosowaniem.
- **Intencja**: Zniechęcanie do oddawania głosów.
- **Szkodliwe**: zakłóca demokratyczne zaangażowanie.

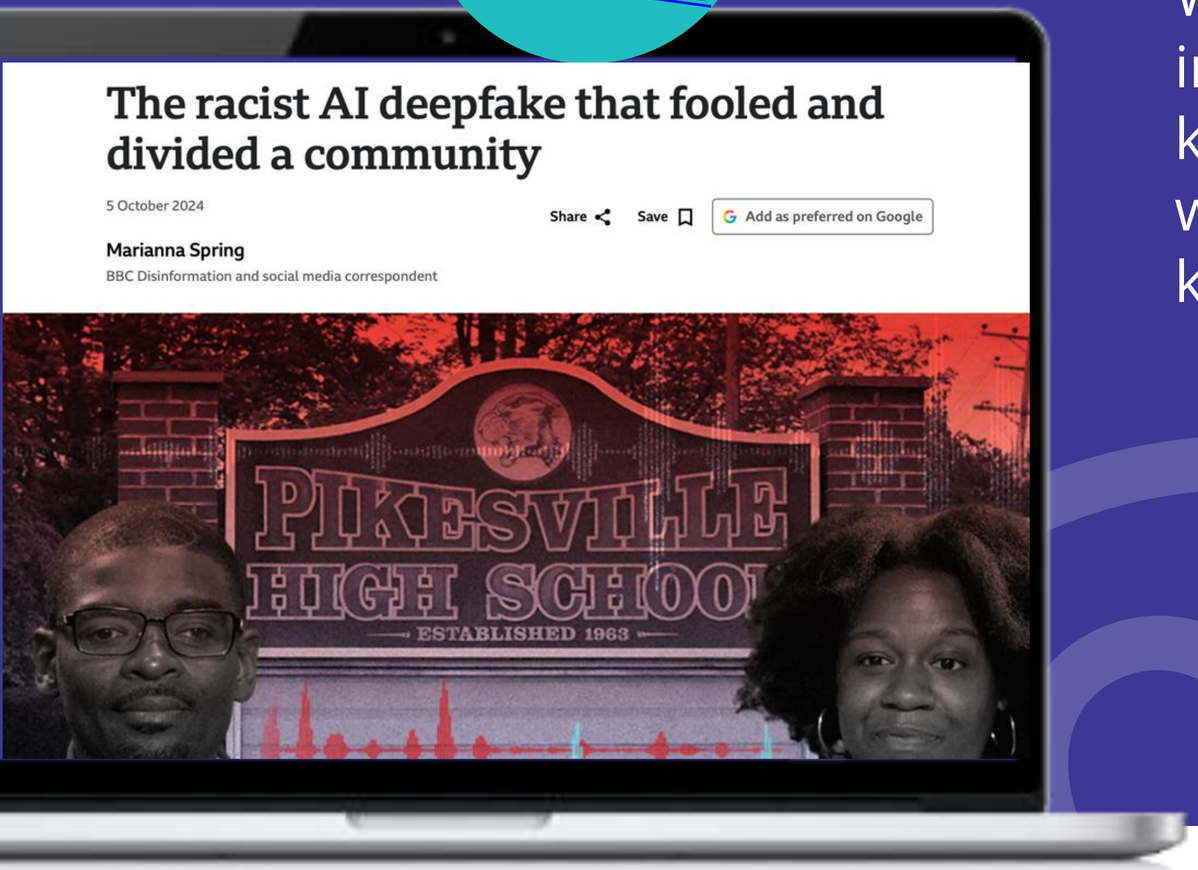
Co powinien uczynić wyborca po odebraniu takiego telefonu?

Ćwiczenie: Odgrywanie ról – dylemat zaawansowanej podróbki

1. Podzielcie się na 3–4 mniejsze grupy.
2. Każda grupa zrealizuje jeden scenariusz.
3. Każdy występ po 2–3 minuty.
4. Refleksja zespołowa.

Studium przypadku „Fałszywe nagranie, które podzieliło społeczność.”

KLIKNIJ, ABY ZOBACZYĆ

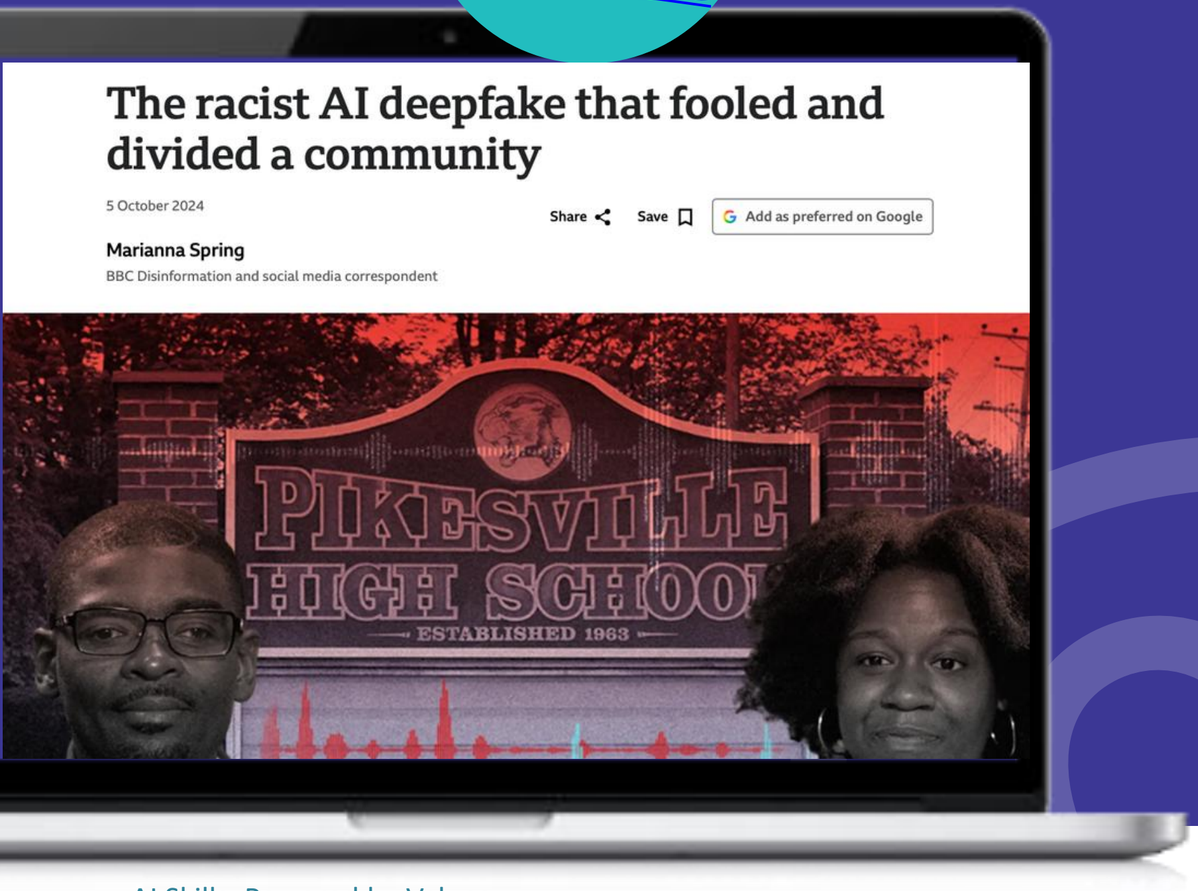


W miasteczku nieopodal Baltimore (USA) w internecie opublikowano nagranie audio, na którym słychać wypowiedzi dyrektora szkoły wygłaszającego rasistowskie i antysemickie komentarze.

- Nagranie błyskawicznie zyskało popularność w mediach społecznościowych.
- Wielu ludzi wierzyło, że to jest prawda.
- Rodzice byli oburzeni.
- Szkoła zaczęła otrzymywać pogróżki.

Studium przypadku „Fałszywe nagranie, które podzieliło społeczność.”

KLIKNIJ, ABY ZOBACZYĆ



Dyrektor szkoły został zawieszony na czas trwania dochodzenia.
Później policja potwierdziła, że nagranie zostało wygenerowane przez sztuczną inteligencję i jest fałszywe. Niemniej jednak, pomimo udowodnienia tego faktu, szkody pozostały; zaufanie do szkoły zostało nadszarpnięte, a ludzie wciąż czuli się zranieni i rozgniewani. Niektórzy nadal wierzyli, że nagranie odzwierciedla „to, co ludzie naprawdę myślą”.

Studium przypadku „Fałszywe nagranie, które podzieliło społeczność”

Dlaczego ludzie w to uwierzyli?
Naukowcy tłumaczą, że ludzie uwierzyli w ten klip, ponieważ:

- Brzmiał autentycznie (klonowanie głosu).
- Zawierał informacje poufne (język instytucji, nazwiska pracowników)
- Dostosowywał się do istniejących obaw i doświadczeń.
- Treści wyłącznie audio zawierają ograniczoną liczbę wskazówek wizualnych.
- Platformy wzmocniły to przed procesem weryfikacji.

Zjawisko to określane jest jako błąd autentyczności: gdy coś wydaje się prawdziwe, przyjmujemy to za prawdę (Vosoughi i in., 2018; Lewandowsky i in., 2012). Sztuczna inteligencja nie tylko wprowadza nas w błąd w zakresie wzroku czy słuchu. Wykorzystuje zaufanie, emocje oraz doświadczenia z przeszłości. Myślenie krytyczne nie polega jedynie na identyfikowaniu błędów technicznych. Chodzi o rozpoznanie, w jaki sposób emocje i przekonania kształtują nasz osąd.

Studium przypadku „Fałszywe nagranie, które podzieliło społeczność”

Co mogło zmniejszyć szkody?

Poproś uczniów o przeprowadzenie dyskusji…

- W którym momencie rozprzestrzenianie się newsa mogłoby zostać zahamowane? *Przed dalszym udostępnieniem? Przed reakcją emocjonalną? Zanim platformy go nagłośnią?*
- Które pytania z metody THINK są tutaj najważniejsze?
  - Prawdziwość? (Czy źródło zostało potwierdzone?)
  - Zamiar? (Kto to przesłał i z jakiego powodu?)
  - Życzliwość? (Kto może doznać krzywdy?)
- Dlaczego stwierdzenie „to wydaje się prawdą” ma większą moc niż „to jest prawdą”?

Podsumowanie dowodów. Twój „Dziennik Detektywistyczny”

- Na tę aktywność przeznaczone jest od 5 do 8 minut, możesz ją wykonać samodzielnie lub w parach.
- Wybierz jeden element zawartości (obraz, wideo, dźwięk lub odpowiedź chatbota).
- Utwórz zwięzłą „Notatkę o wykryciu” według poniższego wzoru.
- Szablon notatki dotyczącej wykrycia (wypełnij)
- Jaka jest zawartość? (1 zdanie)
- Gdzie to odnalazłem? (platforma/konto/link, jeśli to możliwe)
- Moja pierwsza reakcja: (jakie odczucia mnie ogarnęły?)

Podsumowanie dowodów, Twój „Rejestr Śledczy”

Moje wybory (zaznacz to, co wykorzystałeś)

- Weryfikacja szczegółów (co wyglądało nie tak?)
- Sprawdź źródło (kto opublikował? Czy istnieją wiarygodne źródła?)
- Sprawdzenie emocji (czy byłem przymuszany do reakcji?)

Moja decyzja.

- Ufam na tyle, aby go/je wykorzystywać i udostępniać.
- Nie jestem pewien (jeszcze się tym nie podzielę).
- Uważam, że to wprowadza w błąd lub jest nieprawdziwe.

Jedno zdanie uzasadniające (Twoje wyjaśnienie):

Dlaczego to ma znaczenie (krótkie wyjaśnienie): *(Formułując swoje rozumowanie, ukazujesz swoje myślenie, Jest to kluczowy element myślenia krytycznego oraz metapoznania (refleksji nad własnym procesem myślenia)).*

AI Skills, Powered by Values

values

Śledź naszą wędrówkę

www.valuesai.eu

Co-funded by the European Union

Współfinansowane przez Unię Europejską. Wyrażone poglądy i opinie są wyłącznie poglądami i opiniami autora lub autorów i niekoniecznie odzwierciedlają stanowisko Unii Europejskiej ani Fundacji Rozwoju Systemu Edukacji. Ani Unia Europejska, ani instytucja przyznająca grant nie ponoszą za nie odpowiedzialności.