AI Skills,
Powered
by Values
values

Twój przewodnik

Values – przewodnik trenera po modułach szkoleniowvch

www. **valuessai.eu**

Co-funded by
the European Union

ZAWARTOŚĆ

01 Wstęp 3

02 Zestaw startowy VALUES 4

03 Samoocena umiejętności w zakresie sztucznej inteligencji 4

04 Osiem modułów szkoleniowych 5

05 Cele edukacyjne modułów 6

06 Modele zastosowania 16

07 Przydatne linki 17

Co-funded by the European Union

Finansowane przez Unię Europejską. Wyrażone poglądy i opinie są jednak poglądami i opiniami wyłącznie autora/autorów i niekoniecznie odzwierciedlają poglądy Unii Europejskiej ani Fundacji Rozwoju Systemu Edukacji. Ani Unia Europejska, ani Fundacja Rozwoju Systemu Edukacji nie ponoszą za nie odpowiedzialności.

1. Wprowadzenie

Czym jest projekt VALUES?

VALUES - Venturing into AI: Learning for an Unbounded, Ethical, and Sustainable Europe – to projekt partnerski Erasmus+ KA220-YOU Cooperation Partnership in Youth (2024-2026). Projekt, kierowany przez Uniwersytet Szczeciński (Polska), odpowiada na pilną potrzebę: młodzi ludzie, pracownicy młodzieżowi i edukatorzy w całej Europie potrzebują praktycznych, dostępnych i etycznie uzasadnionych zasobów, aby móc korzystać ze sztucznej inteligencji – nie jako bierni użytkownicy, ale jako świadomi obywatele, którzy mogą wpływać na rozwój sztucznej inteligencji.

Projekt wytworzył trzy uzupełniające się produkty, wszystkie dostępne bezpłatnie na stronie www.valuesai.eu:

01 WP2 - Zestaw startowy VALUES dla etycznej sztucznej inteligencji: materiał multimedialny omawiający podstawy etyczne sztucznej inteligencji oparte na siedmiu zasadach UE dotyczących godnej zaufania sztucznej inteligencji.

02 WP3 - Platforma samooceny umiejętności w zakresie sztucznej inteligencji: interaktywne narzędzie online, za pomocą którego uczestnicy oceniają swoją wiedzę na temat sztucznej inteligencji i otrzymują spersonalizowaną ścieżkę kształcenia.

03 WP4 - Osiem modułów OER dotyczących sztucznej inteligencji w edukacji młodzieży: otwarte zasoby edukacyjne, którym towarzyszy ten przewodnik, obejmujące kwestie uczciwości, przejrzystości, dezinformacji, rzetelności, prywatności danych, śladu cyfrowego, legalnego wykorzystania oraz gry decyzyjnej w końcowej fazie projektu.

Czym jest ten Przewodnik Pedagogiczny?

Niniejszy przewodnik jest przeznaczony dla pracowników młodzieżowych, edukatorów, trenerów i wolontariuszy korzystających z ośmiu modułów VALUES. Wyjaśnia on, co zawiera każdy moduł, jak łączy się on z ramami kompetencji DigComp 2.1 i EntreComp oraz jak dostosować go do różnych grup i środowisk. Jest to mapa, a nie scenariusz.

Organizatorzy znajdą: podział wszystkich ośmiu modułów z opisami poszczególnych tematów i rezultatami uczenia się; wskazówki dotyczące integracji Zestawu Startowego (WP2) i Platformy Samooceny (WP3) z modułami szkoleniowymi; sugerowane modele dostarczania (intensywne, cotygodniowe, bez ograniczeń); oraz informacje na temat struktury Odznaki Cyfrowej, która umożliwia uczestnikom formalne potwierdzenie tego, czego się nauczyli.

Ogólne cele edukacyjne

We wszystkich ośmiu modułach uczestnicy będą:

01 Dowiedzą się, czym jest sztuczna inteligencja i dlaczego zasady etyczne mają znaczenie przy jej projektowaniu i użytkowaniu.

02 Będą rozpoznawać i krytycznie oceniać uczciwość, stronniczość, przejrzystość i rozliczalność w systemach sztucznej inteligencji.

03 Zidentyfikują i oprą się dezinformacji, deepfake'om i manipulacjom danymi wykorzystującym sztuczną inteligencję.

04 Zweryfikują treści wspomagane przez sztuczną inteligencję, stosując rygorystyczne metody pozyskiwania informacji, sprawdzania faktów i wykrywania halucynacji.

05 Będą chronić dane osobowe i zarządzać ryzykiem prywatności korzystając z narzędzi AI.

06 Poznają i będą aktywnie zarządzać swoją obecnością cyfrową w środowisku opartym na sztucznej inteligencji.

07 Będą postępować zgodnie z prawem dzięki sztucznej inteligencji – wykorzystując wiedzę na temat praw autorskich, praw osobistych, RODO i unijnej ustawy o sztucznej inteligencji.

08 Zastosują wszystkie osiem obszarów nauki w grze decyzyjnej symulującej rzeczywiste wybory etyczne.

Grupy docelowe

Otwarte zasoby edukacyjne VALUES (VALUES OER) są przeznaczone dla pracowników młodzieżowych i edukatorów pracujących w środowisku pozaformalnym i nieformalnym, a także dla młodych ludzi w wieku 15–30 lat – zwłaszcza tych bezrobotnych, pracujących w niepełnym wymiarze godzin lub oddalonych od edukacji formalnej. Moduły mogą być również wykorzystywane przez uczelnie wyższe, instytucje kształcenia zawodowego, pracodawców i szkoły średnie, które integrują etykę sztucznej inteligencji z istniejącymi programami nauczania.

2. Zestaw startowy VALUES dla etycznej sztucznej inteligencji

Materiały do czytania dla wszystkich moderatorów

Zestaw startowy został opracowany przez EUEI (Dania) we współpracy z Generation (change?) (Malta). Jest on dostępny do bezpłatnego pobrania w języku angielskim, polskim i niemieckim na stronie www.valuesai.eu i składa się z trzech części:

01 Sekcja 1 – Podstawy etyki w dziedzinie sztucznej inteligencji: wprowadzenie do siedmiu zasad UE dotyczących godnej zaufania sztucznej inteligencji, obejmujących m.in. sprawczość człowieka, solidność, prywatność, przejrzystość, różnorodność, dobrostan społeczny i odpowiedzialność.

02 Sekcja 2 – Zastosowania praktyczne: narzędzia ułatwiające pracę, ramy dyskusji i podejścia sesyjne służące integracji etycznej sztucznej inteligencji z istniejącymi programami pracy z młodzieżą.

03 Sekcja 3 — Studia przypadków i referencje: rzeczywiste przykłady zaangażowania młodych ludzi i organizacji w sztuczną inteligencję, zaczerpnięte z krajów partnerskich, mające na celu osadzenie abstrakcyjnych koncepcji w doświadczeniu życiowym.

Animatorzy powinni przepracować Sekcję 1 przed rozpoczęciem któregokolwiek z ośmiu modułów – zawiera ona słownictwo etyczne, do którego odwołuje się cały program. Sekcje 2 i 3 mogą być wykorzystane jako wstęp do sesji, lektura wzbogacająca lub materiały dla uczestników. Cała zawartość Pakietu Startowego jest publikowana na licencji Creative Commons Attribution.

3. Platforma samooceny umiejętności AI

Korzystanie z platformy z grupą

Platforma Samooceny została opracowana przez tvw GmbH (Niemcy) we współpracy z SWTP (Polska). Jest to bezpłatne, interaktywne narzędzie dostępne na stronie www.valuesai.eu. Uczestnicy wypełniają kwestionariusz dotyczący swojej wiedzy z zakresu sztucznej inteligencji i otrzymują spersonalizowany Raport Zwrotny w formacie PDF, który określa ich aktualny poziom, powiązuje wyniki z kompetencjami DigComp i EntreComp oraz rekomenduje moduły, które należy

traktować priorytetowo.

Użyj go jako punktu wyjścia: uczestnicy wypełnią ocenę przed pierwszą sesją, tworząc osobisty punkt odniesienia, który pomoże prowadzącym dostosować program. Możesz również wykorzystać go jako ćwiczenie uzupełniające – wypełniając je przed i po programie – dając uczestnikom mocny wizualny zapis ich nauki do portfolio dowodów w postaci Odznaki Cyfrowej.

01 Ocena trwa 20-30 minut. Uczestnicy potrzebują urządzenia i połączenia internetowego.

02 Raport z opiniami ma charakter osobisty. Należy go traktować jako poufny, chyba że uczestnik zdecyduje się go udostępnić.

03 Uspokój uczestników: nie ma dobrych ani złych odpowiedzi. Narzędzie mapuje ich położenie, a nie testuje ich ani nie ocenia.

4. Osiem modułów szkoleniowych

Przegląd i dostosowanie ram kompetencyjnych

Osiem modułów VALUES stanowi podstawowy materiał dydaktyczny i edukacyjny programu. Każdy z nich zawiera prezentację slajdów, notatki dla moderatorów, ćwiczenia dyskusyjne, zajęcia grupowe i studia przypadków. Wszystkie są dostępne na stronie www.valuesai.eu na licencji Creative Commons, w języku angielskim, polskim i niemieckim.

Moduły te nie są zaprojektowane do realizacji w ustalonej kolejności. Zachęca się moderatorów do korzystania z modułów, które najlepiej pasują do ich grupy, ram czasowych i zagadnień najbardziej istotnych dla uczestników, w dowolnej kolejności, która ma sens w danym kontekście. Jedynym modułem pomyślanym jako naturalny punkt końcowy jest moduł końcowy, który zachęca uczestników do zastosowania przemyśleń z pozostałych modułów – ale nawet ten moduł może zostać wprowadzony wcześniej, jako wstęp, jeśli moderator wyrazi taką wolę.
Moduły są zgodne z dwoma europejskimi ramami kompetencji:

01 DigComp 2.1 (Ramowe ramy kompetencji cyfrowych dla obywateli): pięć obszarów obejmujących: umiejętność korzystania z informacji i danych, komunikację i współpracę, tworzenie treści cyfrowych, bezpieczeństwo oraz rozwiązywanie problemów.

02 EntreComp (Europejskie Ramy Kompetencji Przedsiębiorczości): piętnaście kompetencji obejmujących idee, możliwości, zasoby i działania.

Konkretne kompetencje omówione w każdym module przedstawiono w szczegółowych tabelach zamieszczonych w rozdziale 5 niniejszego przewodnika.

Sprawiedliwość i stronniczość w sztucznej inteligencji | 75 min

Zapewnienie zrównoważonej reprezentacji i punktów widzenia

Przejrzystość w sztucznej inteligencji | 100 min

Bądź transparentny w kwestii treści generowanych przez sztuczną inteligencję

Dezinformacja, deepfake'i i manipulacja danymi | 95 min

Rozpoznawanie i zapobieganie oszustwom opartym na sztucznej inteligencji

Niezawodna treść wspomagana sztuczną inteligencją | 105 min

Źródła, weryfikacja i kontrola halucynacji

Prywatność danych i bezpieczne korzystanie z narzędzi AI | 85 min

Udostępniaj mniej. Chroń więcej. Używaj sztucznej inteligencji odpowiedzialnie.

Sztuczna inteligencja i cyfrowy ślad | 95 min

Kontrolowanie tego, co zostawiasz

Korzystanie ze sztucznej inteligencji zgodnie z prawem| 50 min

Prawa, zgoda i odpowiedzialne korzystanie

Ostateczne podsumowanie: Przeprosiny wspomagane przez sztuczną inteligencję | 60+ min

Zastosowanie wiedzy zdobytej w ramach programu w realistycznym scenariuszu etycznym

|5. Cele edukacyjne modułu (szczegółowe)

Podział tematyczny z mapowaniem ram kompetencji

Poniższe sekcje zawierają szczegółowy opis każdego z ośmiu modułów, w tym kluczowe tematy, efekty uczenia się dla każdego tematu oraz konkretne kompetencje DigComp 2.1 i EntreComp. Tematy są ponumerowane w każdym module wyłącznie dla ułatwienia – same moduły można realizować w dowolnej kolejności, dostosowanej do potrzeb grupy.

|Sprawiedliwość i stronniczość w sztucznej inteligencji

Zapewnienie zrównoważonej reprezentacji i punktów widzenia | 75 minut

Ten moduł przedstawia, jak sztuczna inteligencja może wzmacniać niesprawiedliwość poprzez stronniczość w danych szkoleniowych, decyzjach projektowych lub wdrożeniu. Opierając się na przykładach z życia wziętych – od algorytmów rekrutacyjnych po rozpoznawanie twarzy – uczestnicy zgłębiają znaczenie sprawiedliwości w sztucznej inteligencji, skąd biorą się uprzedzenia i co można z nimi zrobić. Moduł przedstawia uczestnikom pięć praktycznych zasad odpowiedzialnego korzystania ze sztucznej inteligencji.

Temat 1: Wprowadzenie i ważne fakty na temat sztucznej inteligencji

Uczestnicy dowiadują się, że sztuczna inteligencja nie jest neutralna ani nieomylna – odzwierciedla wartości i słabe punkty swoich twórców. Omawiane są kluczowe ograniczenia, w tym skłonność sztucznej inteligencji do generowania zniekształconych lub mylących wyników. Na koniec uczestnicy potrafią wyjaśnić, czym jest sztuczna inteligencja i wymienić trzy jej podstawowe ograniczenia.
DigComp / EntreComp: DigComp 1.1, DigComp 4.1, EntreComp 1.5

Temat 2: Czym jest uczciwość i stronniczość w sztucznej inteligencji i dlaczego to takie ważne?

Przedstawiono podstawowe koncepcje sprawiedliwości i stronniczości algorytmicznej. Uczestnicy zbadali, jak sztuczna inteligencja może traktować grupy w różny sposób – czasami prowadząc do dyskryminacji nawet bez wyraźnego zamiaru. Przykłady wykorzystano w praktyce w różnych sektorach, w tym w zatrudnieniu, scoringu kredytowym i rozpoznawaniu twarzy.
DigComp / EntreComp: DigComp 1.2, EntreComp 1.5, EntreComp 3.4

Temat 3: Pięć zasad uczciwego i odpowiedzialnego korzystania ze sztucznej inteligencji

Uczestnicy poznają pięć praktycznych zasad: wykorzystywanie reprezentatywnych danych; sprawdzanie, czy narzędzie nie jest stronnicze, zanim zaczniesz polegać na nim; zrozumienie działania systemu i jego ograniczeń; pozostawienie kontroli nad decyzjami o wysokim ryzyku przez człowieka; oraz okresowy przegląd systemów AI w miarę zmian danych i kontekstu. Studium przypadku Amazon AI w rekrutacji – w którym narzędzie rekrutacyjne zostało wycofane po tym, jak okazało się, że obniża ono ocenę CV kobiet – ilustruje, dlaczego nadzór ludzki jest tak ważny.
DigComp / EntreComp: DigComp 5.1, EntreComp 1.5, EntreComp 2.1

Temat 4: Źródła błędów w sztucznej inteligencji

Uczestnicy wskazują dwanaście różnych źródeł stronniczości, zilustrowanych rzeczywistym przykładem: stronniczość danych szkoleniowych, stronniczość selekcji, stronniczość etykietowania danych, stronniczość wynikająca z projektowania modelu, stronniczość wynikająca z filtrów bezpieczeństwa i moderacji, stronniczość wynikająca z decyzji dostawcy, stronniczość wynikająca z wersjonowania systemu, stronniczość wynikająca ze sposobu sformułowania pytania, stronniczość interaktywna/konwersacyjna, stronniczość występująca po wdrożeniu/pętli sprzężenia zwrotnego, pomijanie mniej popularnych punktów widzenia oraz stronniczość językowa i kulturowa.
DigComp / EntreComp: DigComp 1.2, DigComp 2.6, EntreComp 1.5

Temat 5: Jak unikać stronniczości?

Dla każdego z dwunastu typów uprzedzeń zidentyfikowanych w Temacie 4 uczestnicy poznają najlepsze praktyki ich redukcji – na przykład wykorzystanie zróżnicowanych i regularnie weryfikowanych danych szkoleniowych, testowanie modeli w grupach przed wdrożeniem, zadawanie neutralnych i precyzyjnych pytań oraz monitorowanie wyników po uruchomieniu. Temat kończy się jedną zasadą przewodnią: traktuj odpowiedź sztucznej inteligencji jako punkt wyjścia, a nie ostateczną prawdę.
DigComp / EntreComp: DigComp 1.2, DigComp 5.3, EntreComp 3.6

Ćwiczenia i aktywności

Samodzielne: ćwiczenie audytowe, w którym uczestnicy identyfikują ukryte uprzedzenia w pozornie neutralnym zadaniu („napisz post na blogu o odnoszących sukcesy przedsiębiorcach”) i przerabiają go, aby był bardziej inkluzywny. Grupowe: uczestnicy porównują, jak różne modele sztucznej inteligencji odpowiadają na ten sam dylemat moralny, a następnie badają, w jaki sposób sztuczna inteligencja generuje treści na podstawie istniejącej pracy ludzkiej (oraz jakie rodzą się z

tym pytania dotyczące uczciwości i autorstwa), ćwiczenie „wykrywanie ukrytych uprzedzeń” porównujące dwa wygenerowane przez sztuczną inteligencję oświadczenia dotyczące przywództwa oraz rzeczywiste dane dotyczące uprzedzeń i stereotypów w dużych modelach językowych.

|Przejrzystość w sztucznej inteligencji

Bądź transparentny w kwestii treści generowanych przez sztuczną inteligencję| 100 minut

W tym module omówiono znaczenie transparentności w kontekście wykorzystania sztucznej inteligencji w tworzeniu treści. Transparentność – jasne ujawnienie, kiedy i jak wykorzystano sztuczną inteligencję – podtrzymuje zaufanie, chroni odpowiedzialność i jest wymagana w wielu kontekstach przez prawo i normy zawodowe. Moduł rozpoczyna się od dyskusji na temat tego, czy kreatywność sztucznej inteligencji różni się od kreatywności człowieka, a następnie przechodzi do implikacji praktycznych i prawnych.

Temat 1: Czym jest przejrzystość i dlaczego jest ważna?

Przejrzystość w sztucznej inteligencji oznacza informowanie odbiorców, kiedy i w jaki sposób sztuczna inteligencja przyczyniła się do powstania treści. Temat bada, dlaczego stawki różnią się w zależności od kontekstu – akademickiego, dziennikarskiego, społecznościowego i opieki zdrowotnej. Ustrukturyzowana dyskusja stawia pytanie, czy każdy wkład sztucznej inteligencji zawsze wymaga ujawnienia.
DigComp / EntreComp: DigComp 2.1, EntreComp 1.5

Temat 2: Treści generowane, wspomagane i edytowane przez sztuczną inteligencję

Wprowadzane są trzy kategorie: w pełni generowane przez sztuczną inteligencję (minimalny wkład człowieka), wspomagane przez sztuczną inteligencję (człowiek wykorzystuje sztuczną inteligencję jako współtwórcę) oraz edytowane przez sztuczną inteligencję (praca człowieka modyfikowana przez sztuczną inteligencję). Uczestnicy ćwiczą kategoryzację rzeczywistych przypadków i rozumieją odpowiedzialność związaną z każdą kategorią.
DigComp / EntreComp: DigComp 3.1, DigComp 1.2, EntreComp 1.4

Temat 3: Przepisy i zasady dotyczące etykietowania treści AI

Wymagania dotyczące etykietowania zawarte w artykule 50 ustawy UE o sztucznej inteligencji są szczegółowo omówione – informowanie użytkowników o interakcji z systemem sztucznej inteligencji, maszynowo czytelne oznaczanie treści syntetycznych oraz obowiązki ujawniania informacji o deepfake'ach i tekstach generowanych przez sztuczną inteligencję w sprawach będących przedmiotem zainteresowania publicznego. Praktyczna zasada: w razie wątpliwości należy ujawnić.
DigComp / EntreComp: DigComp 4.3, EntreComp 1.5

Temat 4: Sztuczna inteligencja, własność intelektualna i autorstwo

Kto jest właścicielem treści generowanych przez sztuczną inteligencję? Uczestnicy dowiedzą się, jak wartość intelektualna, autorstwo i ochrona prawna oddziałują na siebie, gdy w grę wchodzi sztuczna inteligencja, oraz jakie kroki należy podjąć, aby zabezpieczyć autorstwo utworów tworzonych przy pomocy sztucznej inteligencji (dokumentowanie wkładu twórczego, selekcji i decyzji redakcyjnych).
DigComp / EntreComp: DigComp 3.3, EntreComp 1.4, EntreComp 1.5

Temat 5: Proces tworzenia

Uczestnicy uczą się praktycznej, powtarzalnej procedury tworzenia, która pozwoli im odpowiedzialnie korzystać ze sztucznej inteligencji: Twórz (znaj swój cel i odbiorców) → Weryfikacja (sprawdzanie faktów przed zaufaniem płynnemu przekazowi) → Kontrola uczciwości (kto jest uwzględniony, a kto pominięty) → Ujawnianie (po prostu informowanie, kiedy sztuczna inteligencja pomogła). Towarzyszy temu lista „czego unikać" – nie przeinaczaj treści tworzonych przez sztuczną inteligencję, nie kopiuj ich bez przepisywania i podawania źródła, nie wprowadzaj danych osobowych innych osób oraz nie używaj sztucznej inteligencji do atakowania lub stereotypizowania kogokolwiek.
DigComp / EntreComp: DigComp 3.1, EntreComp 1.4

Ćwiczenia i aktywności

Samodzielne: napisz krótkie oświadczenie o przejrzystości dla rzeczywistego lub hipotetycznego dzieła wspomaganego przez sztuczną inteligencję i zaprojektuj zwięzłą etykietę treści AI dla wpisu na blogu, obrazu lub filmu. Grupowe: „Wyzwanie Otwartej Skrzynki", w którym małe grupy analizują scenariusz sztucznej inteligencji typu „black-box" (narzędzie do rekrutacji, system selekcji w szpitalu lub filtr treści w mediach społecznościowych) i wymieniają, co organizacja powinna ujawnić, aby zachować transparentność; oraz ćwiczenie „przed i po" polegające na zastosowaniu procedury „Utwórz-Zweryfikuj-Sprawdź-Ujawnij" do rzeczywistego fragmentu treści.

|Dezinformacja, deepfake'i i manipulacja danymi

Rozpoznawanie i zapobieganie oszustwom opartym na sztucznej inteligencji | 95 minut

Ten moduł przygotowuje uczestników do rozpoznawania i przeciwdziałania oszustwom opartym na sztucznej inteligencji – jednemu z najpilniejszych wyzwań w obecnym środowisku cyfrowym. Obejmuje on dezinformację wspieraną przez sztuczną inteligencję, deepfake'i, praktyczne narzędzia wykrywania, manipulację danymi i kontekstem oraz obowiązki twórców treści. W całym module wykorzystano autentyczne, udokumentowane przypadki, w tym wygenerowany przez sztuczną inteligencję obraz aresztowania Donalda Trumpa z 2023 roku, obraz „Papieża w puchowej kurtce", oszustwo z wykorzystaniem deepfake'a w wideorozmowie w Hongkongu, głos sztucznej inteligencji użyty do podszywania się pod Toma Hanksa w reklamie oraz generowane przez sztuczną inteligencję automatyczne połączenia telefoniczne do Bidena zniechęcające do głosowania.

Temat 1: Dezinformacja: czym jest i jak sztuczna inteligencja ją wzmacnia

Dezinformacja wspierana przez sztuczną inteligencję (AI) jest definiowana jako celowo fałszywe lub wprowadzające w błąd treści tworzone lub rozpowszechniane za pomocą narzędzi AI w celu oszukiwania ludzi i szkodzenia procesom demokratycznym. Studium przypadku zdjęcia aresztowania Trumpa służy do zbadania konsekwencji ekonomicznych, społecznych i demokratycznych. Uczestnicy dyskutują, czy szkody wyrządzone przez dezinformację AI są w pełni odwracalne.
DigComp / EntreComp: DigComp 1.2, DigComp 2.6, EntreComp 1.5

Temat 2: Deepfake'i: czym są i jak je rozpoznać

Uczestnicy dowiedzą się, jak powstają deepfake'i, do czego służą (manipulacja polityczna, oszustwo, niechciane, intymne zdjęcia) i jakie wskaźniki techniczne pomagają je zidentyfikować. Omówione zostaną artefakty wizualne i dźwiękowe charakterystyczne dla sztucznej inteligencji, a także praktyczne ćwiczenia z ich wykrywania.
DigComp / EntreComp: DigComp 1.2, DigComp 4.3, EntreComp 1.5

Temat 3: Manipulacja danymi i kontekstem

Subtelniejsze formy oszustwa wspomaganego przez sztuczną inteligencję są omawiane za pomocą opracowanych

przykładów wizualnych: wybiórczej prezentacji danych, manipulowania skalami wykresów, mylącego skracania osi, sprytnego przegrupowywania kategorii oraz cytatów lub obrazów pozbawionych ich pierwotnego kontekstu.
DigComp / EntreComp: DigComp 1.2, EntreComp 1.5, EntreComp 2.1

Temat 4: Twój zestaw narzędzi detektywa

Uczestnicy otrzymują trzyczęściową listę kontrolną z możliwością podjęcia działań – Sprawdź szczegóły, Sprawdź źródło, Sprawdź intuicję – wraz z dalszymi metodami weryfikacji faktów (odwrotne wyszukiwanie obrazów, sprawdzanie metadanych, odsyłanie do oficjalnych kanałów, geolokalizacja, śledzenie źródeł pierwotnych).
DigComp / EntreComp: DigComp 1.1, DigComp 1.2, DigComp 5.1

Temat 5: Odpowiedzialność i konsekwencje twórców treści

Uczestnicy analizują prawne, społeczne i etyczne konsekwencje tworzenia lub rozpowszechniania dezinformacji, w tym w świetle prawa UE. Internet został zidentyfikowany jako medium o najwyższym ryzyku rozprzestrzeniania dezinformacji, a uczestnicy badają, dlaczego sprostowania rozprzestrzeniają się wolniej niż fałszywe treści.
DigComp / EntreComp: DigComp 3.1, DigComp 4.3, EntreComp 1.5

Ćwiczenia i aktywności

Ćwiczenie polegające na odgrywaniu ról („Dylemat głębokiej fałszerstwa”) oparte na studiach przypadków z życia wziętych, po którym następuje aktywność „Zapis detektywistyczny”, w której uczestnicy stosują zestaw narzędzi detektywistycznych do rzeczywistego fragmentu treści i zapisują swoje rozumowanie, decyzje i poziom pewności siebie.

|Niezawodna treść wspomagana sztuczną inteligencją

Źródła, weryfikacja i kontrola halucynacji | 105 minut

Ten moduł porusza praktyczne wyzwanie tworzenia i oceny rzetelnych treści, gdy sztuczna inteligencja jest częścią procesu twórczego. Wprowadza on halucynacje sztucznej inteligencji – pewne siebie, ale nieprawdziwe wyniki – i zapewnia uczestnikom ramy do weryfikacji, wyszukiwania źródeł i zapewniania jakości pracy wspomaganej przez sztuczną inteligencję. Moduł kończy się krótkim quizem utrwalającym wiedzę.

Temat 1: Tworzenie wiarygodnych treści w erze sztucznej inteligencji

Kluczowe zasady: podawaj źródła, weryfikuj przed publikacją, ujawniaj użycie sztucznej inteligencji, pisz jasno, aktualizuj w przypadku wykrycia błędów, zadbaj o jasne autorstwo, nigdy nie wykorzystuj sztucznej inteligencji do celowego wprowadzania w błąd. Ludzki nadzór jest silniejszy, a nie słabszy, gdy sztuczna inteligencja wspomaga tworzenie treści. Wyjaśniono, co buduje zaufanie odbiorców.
DigComp / EntreComp: DigComp 3.1, DigComp 1.2, EntreComp 1.5

Temat 2: Zagrożenia związane z treściami AI

Przedstawiono trzy kluczowe zagrożenia: ryzyko halucynacji (zmyślone fakty), ryzyko źródła (bezpodstawne lub sfabrykowane cytaty) oraz ryzyko wykrycia (tekst AI jest często rozpoznawalny stylistycznie). Uczestnicy zapoznają się z koncepcją „błędu AI” – treści brzmiącej pewnie, ale niskiej jakości i niedokładnej.
DigComp / EntreComp: DigComp 1.2, DigComp 4.2, EntreComp 1.5

Temat 3: Odpowiedzialny autor

Uczestnicy odpowiadają na cztery pytania dotyczące rozliczalności w odniesieniu do wszystkiego, co stworzą przy pomocy sztucznej inteligencji – dokładność, źródła, oryginalność i rozliczalność – i poznają „listę nigdy nie ufanych" typów treści, które zawsze muszą zostać niezależnie zweryfikowane przed publikacją (cytaty, statystyki, fakty historyczne, definicje).
DigComp / EntreComp: DigComp 1.2, DigComp 5.1, EntreComp 1.5

Temat 4: Praca ze źródłami

Rygorystyczne ramy oceny źródeł: rozróżnianie źródeł pierwotnych od wtórnych, weryfikacja twierdzeń w niezależnych źródłach oraz ocena autora, wydawcy i celu. Praktyczne studium przypadku przedstawia model sztucznej inteligencji, który poprawnie cytuje tytuł artykułu, jednocześnie wymyślając niewłaściwe czasopismo – prawdziwy przykład halucynacji źródłowej – i porównuje nieuzasadnioną odpowiedź sztucznej inteligencji z odpowiedzią wygenerowaną na podstawie rzeczywistej bazy dokumentów.
DigComp / EntreComp: DigComp 1.1, DigComp 1.2, DigComp 5.1

Temat 5: Halucynacje AI

Uczestnicy dowiadują się, czym są halucynacje i dlaczego występują (modele językowe optymalizują sekwencje prawdopodobne, a nie faktyczną dokładność), a także poznają sześć nazwanych typów: halucynacje faktyczne, źródłowe, cytatowe, liczbowe, kontekstowe i mieszane, a także halucynacje „pośrednie", w których imiona i daty wyglądają prawidłowo, ale podstawowe powiązanie z rzeczywistym źródłem jest zerwane.
DigComp / EntreComp: DigComp 1.2, DigComp 5.1, EntreComp 2.1

Temat 6: Weryfikacja treści

Czteroetapowa struktura weryfikacji obejmuje: określenie, które twierdzenia niosą ze sobą największe ryzyko, jeśli okażą się błędne; weryfikację faktów i źródeł na podstawie wiarygodnych, najlepiej podstawowych, źródeł; ocenę jakości, tonu i kontekstu; następnie podjęcie decyzji o opublikowaniu, wstrzymaniu lub odrzuceniu treści.
DigComp / EntreComp: DigComp 1.3, DigComp 5.1, EntreComp 3.2

Temat 7: Analiza danych za pomocą sztucznej inteligencji

Uczestnicy stosują ramy weryfikacji specjalnie do analizy danych wspomaganej przez sztuczną inteligencję, ucząc się rozpoznawać pominięte metody, nieoznaczone obserwacje odstające, nadmiernie uogólnione małe próby oraz błąd potwierdzenia. „Drabina analizy SAFE" oferuje stopniowane, oparte na weryfikacji podejście do udostępniania danych narzędziom AI.
DigComp / EntreComp: DigComp 1.3, DigComp 5.1, EntreComp 3.2

Ćwiczenia i aktywności

Samodzielne: zastosowanie ram weryfikacji do treści wygenerowanej przez sztuczną inteligencję (fragment raportu, twierdzenie statystyczne lub artykuł ze źródłami), identyfikacja wszelkich halucynacji i stworzenie krótkiego raportu z adnotacjami. Grupowe: wspólna weryfikacja faktów złożonego tekstu wspomaganego przez sztuczną inteligencję, z podziałem obowiązków weryfikacyjnych. Kończy się quizem „Bezpieczne dane z AI".

|Prywatność danych i bezpieczne korzystanie z narzędzi AI

Udostępniaj mniej. Chroń więcej. Używaj sztucznej inteligencji odpowiedzialnie. | 85 minut

Ten moduł omawia zagrożenia związane z prywatnością danych, które pojawiają się, gdy osoby i organizacje korzystają z narzędzi AI. Oparty na trzech głównych negatywnych skutkach – narażeniu na utratę prywatności, niesprawiedliwości i fałszywej pewności siebie – moduł oferuje uczestnikom praktyczne wskazówki dotyczące tego, co robić, a czego nie, oraz

ramy wyboru bezpieczniejszych przepływów pracy z wykorzystaniem AI. Moduł omawia, dlaczego bezpieczeństwo danych jest ważne, jak bezpiecznie je udostępniać, prawa do prywatności i zgody oraz możliwości wyboru bezpieczniejszych przepływów pracy.

Temat 1: Dlaczego bezpieczeństwo danych ma znaczenie w przypadku sztucznej inteligencji

Główne zagrożenia: ujawnienie wrażliwych danych (monity wysyłane do sztucznej inteligencji mogą być przechowywane lub pojawiać się w wynikach); ukryta sztuczna inteligencja (niezatwierdzone narzędzia tworzą martwe punkty); naruszenia przepisów prawnych i zgodności; brak przejrzystości dostawców; oraz rozszerzone powierzchnie ataków. Wprowadzono trzy główne negatywne skutki: ryzyko prywatności (ujawnienie), ryzyko uczciwości (szkoda) i ryzyko niezawodności (fałszywe zaufanie).
DigComp / EntreComp: DigComp 4.1, DigComp 4.2, EntreComp 1.5

Temat 2: Jak bezpiecznie udostępniać dane

Praktyczne zasady: unikaj udostępniania danych osobowych, wrażliwych danych osobowych, poufnych informacji organizacyjnych ani danych stron trzecich udostępnianych z zachowaniem poufności. Uczestnicy dowiedzą się, co należy wziąć pod uwagę przed udostępnieniem danych na dowolnej platformie AI – retencja danych, lokalizacja przetwarzania, wykorzystanie danych szkoleniowych – oraz jak anonimizować dane przed udostępnieniem. Zawiera listę kontrolną „Co robić, a czego nie".
DigComp / EntreComp: DigComp 4.2, DigComp 4.1, EntreComp 3.2

Temat 3: Prywatność, prawa i zgoda

Ramy prawne i etyczne: prawa wynikające z RODO (prawo do informacji, prawo do usunięcia danych, prawo do sprzeciwu), wymogi dotyczące zgody na przetwarzanie danych osobowych oraz szczególne obowiązki w przypadku przetwarzania przez sztuczną inteligencję danych wrażliwych, takich jak informacje o stanie zdrowia, dane biometryczne czy dane dotyczące dzieci. Analizowane są rzeczywiste scenariusze dotyczące zgody z udziałem sztucznej inteligencji.
DigComp / EntreComp: DigComp 4.2, DigComp 2.6, EntreComp 1.5

Temat 4: Wybór bezpieczniejszych przepływów pracy w zakresie sztucznej inteligencji

Uczestnicy uczą się oceniać narzędzia AI w kontekście kryteriów prywatności, rozumieją różnicę między usługami AI klasy konsumenckiej i korporacyjnej, korzystają z lokalnych lub działających offline narzędzi AI do zadań wymagających dużej precyzji oraz wdrażają nawyki minimalizacji danych w ramach wspomaganych przez AI przepływów pracy.
DigComp / EntreComp: DigComp 4.1, DigComp 5.1, EntreComp 3.2, EntreComp 2.1

Temat 5: Kradzież danych i poświadczeń – rosnące ryzyko cybernetyczne

Opierając się na najnowszych raportach dotyczących zagrożeń (Zscaler, Microsoft, CrowdStrike, Google Threat Intelligence Group, Anthropic i International AI Safety Report), ten temat obejmuje udokumentowany wzrost liczby ataków ransomware i kradzieży danych uwierzytelniających oraz wstępne dowody na to, że narzędzia AI mogą skutecznie wspierać atakujących. Analizuje również dane z ankiet konsumenckich dotyczące obszarów, w których ludzie czują się najbardziej narażeni – phishing generowany przez sztuczną inteligencję, klonowanie głosu i podszywanie się pod inne osoby, które należą do głównych zagrożeń bezpieczeństwa tożsamości.
DigComp / EntreComp: DigComp 4.1, DigComp 4.2, EntreComp 1.5

Ćwiczenia i aktywności

Samodzielne: audyt osobistego wykorzystania narzędzi AI w ciągu ostatniego tygodnia – jakie dane zostały udostępnione, którym platformom i jakie są zasady dotyczące danych na tych platformach? Sporządź krótką ocenę ryzyka osobistego i zidentyfikuj jedną zmianę w przepływie pracy. Grupowe: oceń zestaw scenariuszy wykorzystania AI w organizacji pod kątem omówionych zasad prywatności, identyfikując zagrożenia i proponując bezpieczniejsze alternatywy.

|Sztuczna inteligencja i cyfrowy ślad

Kontroluj to, co zostawiasz za sobą | 95 minut

Ten moduł pomaga uczestnikom zrozumieć ich cyfrowy ślad – ślad danych pozostawiony przez aktywność online – i łączy go ze sposobem, w jaki systemy sztucznej inteligencji wykorzystują ten ślad, wzmacniają go i są przez niego kształtowane. Uczestnicy dowiadują się, że ich ślad to nie tylko kwestia prywatności: aktywnie wpływa on na decyzje podejmowane przez sztuczną inteligencję. Moduł kładzie nacisk na sprawczość: uczestnicy nie mogą kontrolować wszystkiego, ale mogą podejmować świadome decyzje, które zmniejszają ryzyko dla siebie i innych.

Temat 1: Czym jest cyfrowy ślad?

Rozróżnia się aktywne ślady (świadomie udostępniane informacje) i pasywne ślady (dane gromadzone automatycznie za pomocą plików cookie, adresów IP i sygnałów lokalizacji). Cyfrowe ślady wiążą się z czterema podstawowymi kwestiami: prywatnością, reputacją, bezpieczeństwem i wpływem na innych.
DigComp / EntreComp: DigComp 4.2, DigComp 2.6, EntreComp 1.5

Temat 2: Rodzaje śladów i dlaczego sztuczna inteligencja sprawia, że mają one większe znaczenie

Analizowane są dane behawioralne, dane o lokalizacji, dane z wykresów społecznościowych, dane transakcyjne i dane biometryczne – oraz sposób, w jaki sztuczna inteligencja wykorzystuje je do tworzenia prognoz i zautomatyzowanych decyzji dotyczących uczestników. Uczestnicy analizują decyzje algorytmiczne w swoich codziennych aplikacjach i zastanawiają się nad czterema pytaniami dotyczącymi wartości: Czy dokonuję świadomego wyboru? Czy jestem wykorzystywany jako źródło danych? Czy może to zaszkodzić komuś innemu? Co chciałbym pokazać innym?
DigComp / EntreComp: DigComp 4.2, DigComp 1.3, EntreComp 1.5

Temat 3: Efekt domina: konsekwencje dla Ciebie i innych

Analizowany jest efekt domina cyfrowych śladów – w jaki sposób udostępniane dane o sobie mogą również narażać innych poprzez oznaczanie lokalizacji, udostępnianie zdjęć i zrzuty ekranu. Scenariusze obejmują weryfikację zatrudnienia, ukierunkowane reklamy i profilowanie polityczne. Konsekwencje są często opóźnione i nie do końca odwracalne.
DigComp / EntreComp: DigComp 4.2, DigComp 2.6, EntreComp 3.4

Temat 4: Mój wybór – prosta lista kontrolna i plan działania

Lista kontrolna STOP-THINK-CHECK-CHOOSE pomaga uczestnikom wyrobić sobie powtarzalny nawyk podejmowania decyzji o bezpieczniejszym udostępnianiu treści, stosowany w trzech realistycznych scenariuszach: prośby o uprawnienia aplikacji, oznaczanie innych osób na grupowym zdjęciu i decydowanie, co udostępnić w wrażliwej sytuacji.
DigComp / EntreComp: DigComp 4.2, DigComp 4.1, EntreComp 3.1, EntreComp 2.1

Temat 5: Jak chronić swój cyfrowy ślad

Zestaw konkretnych, praktycznych zasad: sprawdzanie ustawień prywatności, ograniczanie udostępnianych informacji, czyszczenie plików cookie i śledzenie danych, korzystanie z sieci VPN, gdy jest to stosowne, ograniczanie dostępu do informacji o lokalizacji, aktualizowanie oprogramowania antywirusowego, monitorowanie informacji finansowych i medycznych, unikanie nieznanych linków i aktualizowanie haseł po naruszeniu.
DigComp / EntreComp: DigComp 4.2, DigComp 4.1, EntreComp 3.1, EntreComp 2.1

Ćwiczenia i aktywności

Moduł kończy się ustrukturyzowanym audytem, quizem i autotestem (20 minut): quiz sprawdzający zrozumienie kluczowych pojęć, samoocena „sprawdź swój ślad cyfrowy" z wynikami podzielonymi na poziomy oraz audyt kierowany, w którym uczestnicy identyfikują trzy ryzyka związane ze śladem cyfrowym na używanej przez siebie platformie, dopasowują każde z nich do poprawy i zobowiązują się do podjęcia jednego działania dzisiaj, jednego w tym tygodniu i jednego stałego nawyku.

|Działanie zgodnie z prawem przy użyciu sztucznej inteligencji

Prawa, zgoda i odpowiedzialne korzystanie | 50 minut

Ten moduł porusza kwestie prawne związane z wykorzystaniem sztucznej inteligencji (AI) – obszaru, w którym wielu młodych ludzi ma ograniczoną wiedzę, ale jest narażonych na znaczne ryzyko. Moduł porządkuje krajobraz prawny wokół trzech podstawowych poziomów ryzyka (prawa osobiste, prawa autorskie i regulacje) i udostępnia uczestnikom praktyczne narzędzia do sprawdzenia przed tworzeniem lub udostępnianiem treści wspomaganych przez AI.

Temat 1: Dlaczego ważne jest, aby działać zgodnie z prawem w kontekście sztucznej inteligencji

To, że sztuczna inteligencja coś umożliwia, nie oznacza, że jest to zgodne z prawem. Narzędzia AI potrafią klonować głosy, generować twarze i ponownie wykorzystywać tekst w ciągu kilku sekund – ale konsekwencje prawne, w tym grzywny, pozwy sądowe i usuwanie treści, ponoszą osoby, które wykorzystują lub publikują dane wyjściowe, a nie same narzędzia AI. Wprowadzono trzy poziomy ryzyka prawnego: ryzyko naruszenia praw osobistych, ryzyko naruszenia praw autorskich i ryzyko regulacyjne.
DigComp / EntreComp: DigComp 4.3, EntreComp 1.5

Temat 2: Prawa związane z głosem i sztuczna inteligencja

Głos danej osoby podlega ochronie prawnej – jako dane osobowe i potencjalnie biometryczne zgodnie z RODO, jako kwestia osobowości i prawa do wizerunku oraz jako prawo wykonawcy w istniejących nagraniach. „Reguła zgody na wykorzystanie głosu" zawiera trzy pytania do odpowiedzi przed klonowaniem lub imitacją prawdziwego głosu.
DigComp / EntreComp: DigComp 4.3, DigComp 4.2, EntreComp 1.5

Temat 3: Prawa do wizerunku i podobieństwa

W przypadku wykorzystania sztucznej inteligencji z wizerunkiem prawdziwej osoby obowiązują dwie warstwy prawne: prawa osoby przedstawionej na zdjęciu i prawa fotografa. Wprowadzono „Listę kontrolną wykorzystania wizerunku" (prawo do zdjęcia, prawo osoby, kontekst, ujawnienie) jako praktyczne narzędzie.
DigComp / EntreComp: DigComp 4.3, DigComp 4.2, EntreComp 1.5

Temat 4: Prawa autorskie w erze sztucznej inteligencji

Co się dzieje, gdy sztuczna inteligencja jest trenowana na utworach chronionych prawem autorskim bez zezwolenia; czy wyniki sztucznej inteligencji mogą być chronione prawem autorskim i kto może je uzyskać; jak unikać naruszeń praw autorskich; oraz jakie materiały objęte licencją Creative Commons i domeną publiczną można swobodnie wykorzystywać.

Przedstawiono praktyczne narzędzie do sprawdzania praw autorskich.
DigComp / EntreComp: DigComp 3.3, EntreComp 1.4, EntreComp 1.5

Temat 5: Prawo dotyczące sztucznej inteligencji: Ustawa UE o sztucznej inteligencji i RODO

Przystępny przegląd dwóch najważniejszych unijnych ram prawnych dotyczących sztucznej inteligencji: RODO (prawa do ochrony danych i prywatności) oraz unijnej ustawy o sztucznej inteligencji (kategoryzacja ryzyka, niedozwolone praktyki, obowiązki w zakresie przejrzystości). Skupiono się na przepisach najbardziej istotnych dla codziennego korzystania ze sztucznej inteligencji przez młodych ludzi – wystarczająco, aby wiedzieć, kiedy szukać porady.
DigComp / EntreComp: DigComp 4.2, EntreComp 1.5

Ćwiczenia i aktywności

Ćwiczenia konsolidują trzy narzędzia kontrolne (Voice Gate, Image Checklist, Copyright Check). Samodzielne: zastosuj wszystkie trzy narzędzia w pięciu scenariuszach wykorzystania sztucznej inteligencji, identyfikując ryzyko i kroki przed przystąpieniem do działania. Grupowe: złożony scenariusz (organizacja młodzieżowa tworząca materiały promocyjne generowane przez sztuczną inteligencję z udziałem prawdziwych uczestników) wymagający pełnej, trzypoziomowej oceny ryzyka prawnego z zalecanymi środkami zaradczymi.

|Moduł finałowy: wiadomość z przeprosinami wspomagana przez sztuczną inteligencję

Zastosowanie wiedzy zdobytej w ramach programu w realistycznym scenariuszu etycznym | 60+ minut

Zamiast wprowadzać nowe treści, ten punkt kulminacyjny umieszcza uczestników w realistycznym scenariuszu – w wiadomości z przeprosinami wspomaganej przez sztuczną inteligencję – który czerpie z przemyśleń z całego programu. Sztuczna inteligencja wkroczyła w nasze najbardziej intymne rozmowy, tworząc przeprosiny, wiadomości o rozstaniu i kondolencje. To rodzi pytania o autentyczność, odpowiedzialność i naprawę relacji, które wykraczają poza wszystkie zasady etyczne poruszane w programie.

Temat 1: Tło: Dlaczego sztuczna inteligencja ma znaczenie w komunikacji osobistej

Kontekst i stawka scenariusza: Sztuczna inteligencja przeszła od sprawdzania pisowni e-maili do tworzenia wiadomości o znaczeniu emocjonalnym. Zidentyfikowano trzy zagrożenia: szybkość (sztuczna inteligencja tworzy wiadomości w kilka sekund; naprawa zajmuje czas), płynność (płynne sformułowania mogą ukryć winę lub błędy rzeczowe) oraz dystans (bezpieczniej jest pozwolić maszynie mówić, ale druga osoba nadal odbiera słowa).
DigComp / EntreComp: EntreComp 1.5, EntreComp 2.1

Temat 2: Anatomia szczerych przeprosin

Opierając się na badaniach nad konfliktami (Lazare; Lewicki), uczestnicy uczą się czteroelementowego schematu: Przyznaj się (jasno nazwij, co się stało); Przyjmij odpowiedzialność (przyjmij swoją część, unikaj zrzucania winy na innych); Napraw (konkretnie zajmij się skutkami); Zmień (opisz, co się zmieni, obiecuj tylko to, czego możesz dotrzymać). Wgląd: krótka, szczera wiadomość obejmująca wszystkie cztery ruchy zwykle jest skuteczniejsza niż długa, dopracowana wiadomość, która pomija którykolwiek z nich.

DigComp / EntreComp: EntreComp 1.5, EntreComp 3.4, EntreComp 3.6

Temat 3: Gdzie sztuczna inteligencja pomaga, a gdzie zawodzi

Modele językowe są mocne pod względem formy (struktura, rejestr, opcje frazowania, dostępność dla osób niebędących rodzimymi użytkownikami języka), ale słabe pod względem prawdy (fakty, przerzucanie winy, składanie zbyt wygórowanych obietnic, prywatność). Cztery zabezpieczenia – Weryfikacja, Uczciwość, Prywatność i Ujawnianie – dają uczestnikom możliwość sprawdzenia przed wysłaniem jakiejkolwiek wiadomości wspomaganej przez sztuczną inteligencję.
DigComp / EntreComp: DigComp 3.1, EntreComp 1.5, EntreComp 2.1

Temat 4: Gra decyzyjna

Główna aktywność. Uczestnicy czytają scenariusz (konflikt, który przeniósł się na czat grupowy), wybierają ścieżkę – korzystając ze sztucznej inteligencji i wysyłając go w obecnej formie, tworząc szkic z pomocą sztucznej inteligencji, a następnie go edytując, lub pisząc go samodzielnie – stosują co najmniej dwa z promptów z zestawu narzędzi (STOP-THINK-CHECK-CHOOSE, THINK, Fairness Check, Verification) i podążają za swoim wyborem do jednego z kilku rozgałęzionych rezultatów, z których każdy zawiera otwarte pytania do dyskusji.
DigComp / EntreComp: EntreComp 1.5, EntreComp 3.4, EntreComp 3.6, DigComp 3.1

Temat 5: Gromadzenie dowodów

Uczestnicy rejestrują swoją ścieżkę, narzędzia, których używali, co zweryfikowali, kogo mogło to dotyczyć, oraz co ujawniliby na temat roli sztucznej inteligencji – rejestrując to bezpośrednio w portfolio dowodów Digital Badge.
DigComp / EntreComp: EntreComp 1.5, EntreComp 3.6

Ćwiczenia i aktywności

Ten moduł ma przede wszystkim charakter partycypacyjny: małe grupy biorą udział w grze decyzyjnej, zapisują rozumowanie w każdym punkcie decyzyjnym i przedstawiają wybory grupie, a następnie prowadzone jest moderowane omówienie łączące wybory z ramami omawianymi w innych częściach programu.

6. Modele zastosowania

Dostosowywanie programu do kontekstu

Intensywny (2-3 dni)

Wszystkie osiem modułów w skoncentrowanej formie, z podsumowaniem zamykającym program po zapoznaniu się z pozostałymi modułami. Nadaje się do pobytów studenckich, wymian młodzieżowych, mobilności w ramach programu Erasmus+ oraz dni rozwoju kadry.

Sesje tygodniowe (8-10 tygodni)

Jeden moduł tygodniowo, po 60–105 minut na sesję. Sesja wprowadzająca do Platformy Samooceny oraz sesja zamykająca portfolio Odznak Cyfrowych tworzą kompletny, 10-tygodniowy program. Program odpowiedni dla ośrodków młodzieżowych i organizacji z regularnymi, cotygodniowymi grupami.

Modułowy / Tematyczny

Moduły wybrane w oparciu o potrzeby grupy, niekoniecznie obejmujące wszystkie osiem. Moduły dotyczące uczciwości, przejrzystości, dezinformacji i rzetelnych treści dobrze sprawdzają się jako samodzielny dział poświęcony treściom i etyce

AI; moduły dotyczące prywatności danych, śladu cyfrowego i legalnego działania tworzą kolejną grupę. Moduł finałowy stanowi samodzielne ćwiczenie zamykające dowolną kombinację pozostałych modułów.

Mieszane / Hybrydowe

Niektóre moduły realizowane są stacjonarnie, inne online lub w formie samodzielnej nauki. Zestaw startowy i platforma samooceny idealnie nadają się do samodzielnej pracy online między sesjami. Każdy moduł można podzielić na samodzielną powtórkę treści i sesję dyskusji moderowanej.

W pełni online

Wszystkie moduły za pośrednictwem wideokonferencji. Sale spotkań umożliwiają zajęcia w małych grupach; wspólne tablice cyfrowe (Miro, Menti) umożliwiają ćwiczenia grupowe. Wskazówki dotyczące prowadzenia zajęć online znajdują się w notatkach dla prowadzącego do każdego modułu.

Edukacja rówieśnicza

Młodzi ludzie, którzy ukończyli program, są szkoleni jako edukatorzy rówieśniczy, którzy prowadzą moduły w nowych grupach. Model ten znacząco poszerza zasięg i pogłębia wiedzę edukatorów rówieśniczych. Gra decyzyjna w ramach kursu końcowego jest szczególnie skuteczna w formacie edukacji rówieśniczej.

Konteksty implementacji

01 Model otwarty: moduły wykorzystywane jako samodzielne lub sekwencyjne zasoby na zasadzie dobrowolności – najczęściej spotykane w nieformalnej pracy z młodzieżą.

02 Model obowiązkowy: program VALUES włączony do istniejącego programu nauczania lub kwalifikacji, przy czym mechanizmem oceny jest Odznaka Cyfrowa.

03 Model pozalekcyjny: zajęcia fakultatywne w ramach placówki edukacyjnej, programu szkoleniowego pracodawcy lub organizacji społecznej.

04 Kształcenie ustawiczne/uczenie się przez całe życie: moduły wykorzystywane w ramach CPD, a Odznaka Cyfrowa zapewnia formalne uznanie.

05 Formalne przyjęcie (HEI/VET): poszczególne moduły przyjęte w ramach istniejących kursów lub cały program jako samodzielny przedmiot fakultatywny. Powiązanie DigComp i EntreComp stanowi argument za równoważnością akademicką na rzecz uznania instytucjonalnego.

7. Przydatne linki

Zasoby dla moderatorów i uczestników

Strona internetowa projektu VALUES

www.valuesai.eu – wszystkie osiem modułów, Zestaw Startowy i Platforma Samooceny. Zasoby pozostają dostępne przez co najmniej pięć lat po zakończeniu projektu.

WP2 – Zestaw startowy VALUES dla etycznej sztucznej inteligencji

Dostępny bezpłatnie na stronie www.valuesai.eu w języku angielskim, polskim i niemieckim.

WP3 – Platforma samooceny umiejętności AI

Dostępna bezpłatnie na stronie www.valuesai.eu. Generuje spersonalizowany Raport Zwrotny w formacie PDF. Zalecane jako aktywność przed i po programie.

WP4 – Wszystkie osiem modułów OER

Wszystkie slajdy i przewodnik trenera dostępne na stronie www.valuesai.eu. Licencja Creative Commons – możliwość dowolnej adaptacji z podaniem autorstwa.

Odznaki cyfrowe VALUES

Opisy kompetencji, wytyczne dotyczące portfolio dowodów oraz możliwość składania odznak dostępne są na stronie www.valuesai.eu. W razie pytań prosimy o kontakt z Momentum (liderem WP4: www.momentumconsulting.ie).

Struktura DigComp 2.1

https://joint-research-centre.ec.europa.eu/digcomp_pl

Framework EntreComp

https://joint-research-centre.ec.europa.eu/entrecomp-entrepreneurship-competence-framework_en

Ustawa UE o sztucznej inteligencji

https://artificialintelligenceact.eu - dotyczy modułów dotyczących prywatności danych, śladu cyfrowego i prawnego wykorzystania sztucznej inteligencji.

Wytyczne etyczne UE dotyczące godnej zaufania sztucznej inteligencji

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai - siedem zasad stanowiących podstawę wszystkich ośmiu modułów.

Tekst RODO

https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679 - dotyczy modułów dotyczących prywatności danych, śladu cyfrowego i prawnego wykorzystania sztucznej inteligencji.

AI Skills, Powered by Values

values

www.valuesai.eu

MODUŁ 01

Sprawiedliwość i stronniczość w sztucznej inteligencji

Zapewnienie zrównoważonej reprezentacji oraz różnorodnych perspektyw

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

01	Wprowadzenie oraz kluczowe informacje dotyczące sztucznej inteligencji (10 min)	3
02	Czym jest uczciwość i stronniczość w sztucznej inteligencji oraz dlaczego ma to kluczowe znaczenie? (10 min)	8
03	Jak właściwie korzystać ze sztucznej inteligencji? Pięć zasad (15 min)	11
04	Źródła stronniczości w sztucznej inteligencji (20 min)	25
05	Jak unikać stronniczości? (10 min)	40
06	Przykłady, ćwiczenia, informacje i wskazówki (20 min)	45

Materiał dostępny na licencji Creative CommonsCC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

by VALUES

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

Wprowadzenie oraz kluczowe informacje dotyczące sztucznej inteligencji

(10 minut)

Sztuczna inteligencja jest szeroko stosowana w tworzeniu treści cyfrowych, takich jak teksty, obrazy i filmy.

Treści te kształtują sposób myślenia i podejmowania decyzji przez ludzi. Z tego względu istotne jest zrozumienie, w jaki sposób wykorzystywana jest sztuczna inteligencja oraz jak zapewnić uczciwość i odpowiedzialność w treściach generowanych przez nią.

AI Skills,
Powered
by Values
values

FAKT 1

Sztuczna inteligencja nie jest nieomylna. Bazuje na danych, które nie są weryfikowane pod kątem ich dokładności, co może prowadzić do mylnych wniosków.

„Ludzie sądzą, że mogą ufać informacjom dostarczanym przez asystentów AI, jednak badania wskazują, że potrafią oni generować odpowiedzi na pytania dotyczące istotnych wydarzeń, które są zniekształcone, niezgodne z faktami lub wprowadzające w błąd.”

Pete Archer, dyrektor programowy ds. generatywnej sztucznej inteligencji w BBC, 11 lutego 2025 r.

FAKT 2

Sztuczna inteligencja to forma inteligencji stworzona przez człowieka. Opiera się na algorytmach opracowanych przez ludzi. Algorytmy te różnią się czułością, dokładnością, jakością oraz innymi parametrami, które mają istotny wpływ na generowane informacje zwrotne.

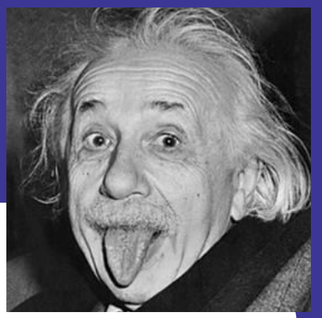
„To, czego uczymy sztuczną inteligencję, odsłania nasze własne wartości.”

„Musimy upewnić się, że systemy sztucznej inteligencji są zgodne z wartościami ludzkimi.”

Demis Hassabis, dyrektor generalny Google DeepMind.

FAKT 3

Zwróć uwagę na sposób, w jaki formułowane są pytania.
Wiele zależy od dokładności sformułowań oraz poziomu Twojej wiedzy.



„Gdybym miał godzinę na rozwiązanie problemu, a moje życie zależałoby od wyniku, poświęciłbym pierwsze 55 minut na sformułowanie właściwego pytania. Gdybym je znał, mógłbym rozwiązać problem w mniej niż pięć minut.”

Albert Einstein

02

Czym jest uczciwość i stronniczość w sztucznej inteligencji oraz dlaczego ma to kluczowe znaczenie?

(10 minut)

Czym jest stronniczość?

Błędy w sztucznej inteligencji mogą występować na różnych etapach: od gromadzenia danych, przez projektowanie modelu, aż po wdrażanie.

Uprzedzenia oznaczają niesprawiedliwość lub preferowanie niektórych osób kosztem innych.
W kontekście sztucznej inteligencji stronniczość występuje, gdy system jest szkolony na ograniczonej lub niezrównoważonej liczbie danych, co prowadzi do powielania tych wzorców.

W sztucznej inteligencji stronniczość może przybierać następujące formy:

- Niektóre osoby są prezentowane częściej niż inne.
- Niektóre grupy są pomijane.
- Stereotypy się powtarzają.

Dlaczego stronniczość jest istotna:

- Sztuczna inteligencja ma potencjał oddziaływania na wiele osób jednocześnie.
- Niesprawiedliwe wyniki mogą prowadzić do rzeczywistych szkód.
- Ludzie powinni weryfikować i korygować wyniki sztucznej inteligencji.

Dlaczego sprawiedliwość ma znaczenie?

- Sprawiedliwe traktowanie osób i grup społecznych
- Unikanie uprzedzeń oraz dyskryminacji
- Równa reprezentacja i integracja
- Odpowiedzialność jednostki za etyczne wykorzystanie sztucznej inteligencji

„Sztuczna inteligencja będzie albo największym osiągnięciem, albo najpoważniejszym zagrożeniem, jakie kiedykolwiek spotka ludzkość.”

Elon Musk

03

Jak właściwie korzystać ze sztucznej inteligencji? Pięć zasad

(15 minut)

5 zasad właściwego korzystania ze sztucznej inteligencji

01 Używaj rzetelnych i uczciwych danych.

02 Sprawdzaj nieuczciwe schematy

03 Wyjaśnij jak coś funkcjonuje

04 Utrzymuj zaangażowanie ludzi

05 Regularnie sprawdzaj wyniki.

Zasada 01

Sztuczna inteligencja uczy się na podstawie informacji, które jej dostarczamy. Jeśli te informacje nie obejmują pewnych grup ludzi, sztuczna inteligencja nie będzie funkcjonować efektywnie dla wszystkich. Aby zapewnić sprawiedliwość, musimy korzystać z danych, które równomiernie reprezentują wszystkich.

Globalna ilość danych w Internecie wynosi obecnie około 200 zettabajtów (ZB) i rośnie w sposób wykładniczy. Jeden zettabajt to jeden bilion gigabajtów (GB).

1 ZB = 1 000 000 000 000 GB

Jednak nie wszystkie te informacje są **rzetelne, aktualne i wiarygodne.**

Użyj potwierdzonych danych.

Jeżeli jakaś informacja jest dla Ciebie szczególnie ważna, warto zweryfikować ją u źródła.

- Pamiętaj: niezależnie od źródła danych, zawsze dokonuj ich własnej krytycznej analizy.
- Bądź uważny na sygnały dyskryminacji oraz nierównego traktowania.

Przykłady rzetelnych źródeł danych i informacji

- Oficjalne strony internetowe rządowe lub uznawane instytucje międzynarodowe.
- Raporty tematyczne opracowywane przez uznawane instytucje, takie jak Organizacja Narodów Zjednoczonych oraz jej wyspecjalizowane agencje, w tym UNESCO, FAO i ILO.
- Bazy danych zawierające oficjalne dane statystyczne są prowadzone przez uznane organizacje, takie jak Eurostat — oficjalna baza danych statystycznych Unii Europejskiej.
- Uznane bazy danych zawierające publikacje naukowe, takie jak Scopus, Web of Science, PubMed oraz ScienceDirect.

Zasada 02

Zanim rozpoczniemy korzystanie z narzędzia AI, musimy je przetestować. Powinniśmy poszukiwać „stronniczości” – sytuacji, w której AI traktuje jedną grupę ludzi lepiej niż inną. W przypadku wykrycia problemu, należy go niezwłocznie rozwiązać.

Testowanie narzędzi i systemów w kontekście stronniczości polega na weryfikacji, czy ich wyniki nie są systematycznie zniekształcane na korzyść lub niekorzyść określonych grup, tematów lub scenariuszy.

Innymi słowy, dotyczy to identyfikacji niesprawiedliwości w danych, modelu, regułach, które go regulują, oraz w sposobie jego zastosowania.

Jak wygląda takie testowanie w praktyce?

- Porównanie odpowiedzi systemu dla różnych grup użytkowników lub profili.
- Sprawdzenie, czy model preferuje jedną płeć, narodowość, grupę wiekową, język lub status społeczny
- Analiza danych szkoleniowych w kontekście braku reprezentatywności
- Testowanie z wykorzystaniem kontrolowanych zestawów podobnych pytań, w których modyfikowane są jedynie wrażliwe atrybuty.
- Ocena, czy system reprodukuje stereotypy lub wytwarza niesprawiedliwe rekomendacje.

Zasada 03

Aby uzyskać dokładne wyniki od sztucznej inteligencji, należy zrozumieć, jak funkcjonuje system, na jakich zasadach się opiera oraz jakie są jego ograniczenia.

Sztuczna inteligencja funkcjonuje na podstawie danych, wzorców, instrukcji oraz reguł optymalizacji, dlatego jakość jej odpowiedzi jest uzależniona od jakości danych wejściowych, kontekstu oraz sposobu wykorzystania systemu.

Kluczowe zasady

- Podaj kontekst, cel oraz odbiorców.
- Zamiast nieprecyzyjnych sformułowań, korzystaj z konkretnych informacji.
- Sprawdź, czy odpowiedź jest zgodna z rzetelnymi źródłami.
- Pamiętajmy, że sztuczna inteligencja nie „wie” w ludzkim sensie tego słowa; przetwarza wzorce na podstawie danych i instrukcji.

Zasada 04

Nie powinniśmy pozwalać sztucznej inteligencji na podejmowanie wszystkich istotnych decyzji. Ludzie zawsze powinni sprawować nadzór nad sztuczną inteligencją.

Ostateczna decyzja powinna należeć do człowieka, szczególnie w tak istotnych sprawach jak praca czy zdrowie, aby mieć pewność, że sztuczna inteligencja jest dla niego przyjazna i sprawiedliwa.

„Musimy upewnić się, że sztuczna inteligencja pozostanie pod kontrolą człowieka.”

Profesor Uniwersytetu w Montrealu, ekspert i badacz systemów sztucznej inteligencji.

Koncepcja zaangażowania człowieka w procesy technologiczne, szczególnie w aspekcie decyzyjnym, stanowi jedno z kluczowych założeń strategii UE: Przemysł 5.0. Jej celem jest ukierunkowanie oraz zrównoważenie rozwoju technologicznego w sposób, który przynosi maksymalne korzyści i wspiera ochronę wartości ludzkich, zamiast je naruszać.

AI Skills,
Powered
by Values
values

STUDIUM PRZYPADKU:

Narzędzie rekrutacyjne Amazon AI

- Amazon stworzył system rekrutacyjny oparty na sztucznej inteligencji, który służy do wstępnej selekcji kandydatów do pracy. Niestety, system ten wyciągnął wnioski z wcześniejszych schematów zatrudniania i zaczął obniżać oceny CV, które zawierały elementy kojarzone z kobietami.
- Ponieważ problem został zidentyfikowany przez recenzentów systemu, Amazon zrezygnował z narzędzia przed jego pełnym wdrożeniem.
- To wyraźny przykład, w którym ludzki nadzór zapobiegł przekształceniu błędu sztucznej inteligencji w decyzję operacyjną, szczególnie w tak ryzykownym obszarze jak rekrutacja.

Zasada 05

Świat ulega zmianom, sztuczna inteligencja również. Powinniśmy regularnie oceniać sztuczną inteligencję co kilka miesięcy, aby upewnić się, że funkcjonuje prawidłowo. Jeśli zacznie popełniać błędy lub wykazywać nieuczciwe zachowania, konieczna będzie jej aktualizacja.

Dlaczego to jest istotne

Sztuczna inteligencja nie jest doskonała po wdrożeniu, ponieważ dane ją otaczające zmieniają się w czasie. Zachowania użytkowników, warunki społeczne, zasady biznesowe, a nawet język mogą ulegać zmianom, co może sprawić, że starszy model stanie się mniej wiarygodny.

Regularne przeglądy zazwyczaj koncentrują się na wydajności, stronniczości oraz odchyleniu danych.

Zespoły analizują aktualne wyniki w kontekście wcześniejszych testów; oceniają, czy niektóre grupy są traktowane niesprawiedliwie; a także podejmują decyzje dotyczące konieczności ponownego szkolenia lub dostosowania modelu.

„Ufaj, lecz weryfikuj.”
Ronald Reagan

04

Źródła stronniczości w sztucznej inteligencji

(20 minut)

Błąd w systemach sztucznej inteligencji

01

Błąd w danych szkoleniowych

02

Błąd w selekcji

03

Błąd w etykietowaniu danych

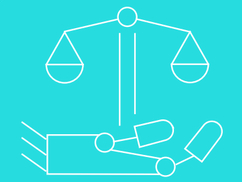
04

Błąd wynikający z konstrukcji modelu

05

Błąd spowodowany filtrami bezpieczeństwa i moderacji

06



Błąd wynikający z decyzji dostawców narzędzi sztucznej inteligencji

07

Błąd spowodowany wersją systemu AI

08

Błąd wynikający z formuły, w jakiej użytkownik zadaje pytanie.

09

Błąd interaktywny / konwersacyjny

10

Błąd po wdrożeniu / Błąd pętli sprzężenia zwrotnego

11

Błąd wynikający z ignorowania mniej popularnych perspektyw

12

Uprzedzenia językowe oraz kulturowe

01 Błąd w danych szkoleniowych

Błąd w danych szkoleniowych odnosi się do nieprawidłowości w zbiorze danych wykorzystanym do trenowania modelu sztucznej inteligencji. Oznacza to, że zbiór danych może być niezrównoważony, niereprezentatywny lub obciążony uprzedzeniami, co może prowadzić do tego, że model nauczy się zniekształconych wzorców.

PRZYKŁAD:

Przykładem może być system rozpoznawania twarzy, który jest głównie trenowany na zdjęciach osób o jasnej karnacji. Taki model może mieć trudności z rozpoznawaniem osób o ciemniejszej karnacji, ponieważ w danych treningowych występowały ich zbyt małe ilości.

02 Błąd w selekcji

Błąd selekcji to błąd próby, który występuje, gdy osoby lub dane uwzględnione w analizie nie są reprezentatywne dla całej populacji, co może prowadzić do zniekształcenia wyników.

PRZYKŁAD:

Jeśli analizujesz jakość usług aplikacji jedynie wśród najbardziej aktywnych użytkowników, wyniki mogą być zbyt optymistyczne, ponieważ pomijasz klientów mniej zadowolonych lub mniej aktywnych.

03

Błąd w etykietowaniu danych

Błąd w etykietowaniu danych występuje, gdy osoby lub narzędzia przypisujące etykiety danym działają subiektywnie, niespójnie lub z uprzedzeniami.

PRZYKŁAD:

Jeśli te same obrazy zostaną oznaczone przez niektórych autorów jako „agresywne”, a przez innych jako „neutralne”, model uczy się zniekształconych schematów.

04

Błąd wynikający z konstrukcji modelu

Błąd wynikający z projektu modelu to błąd spowodowany architekturą oraz decyzjami projektowymi samego modelu, w tym stosowanymi funkcjami i metodami optymalizacji decyzji.

PRZYKŁAD:

System rekrutacji został zaprojektowany w sposób, który faworyzuje kandydatów o określonym stylu CV lub ścieżce kariery, podobnej do osób już zatrudnionych. W rezultacie może on obniżyć ocenę wysoko wykwalifikowanych kandydatów jedynie z powodu tego, że ich doświadczenie zawodowe wydaje się mniej „standardowe”.

05

Błąd spowodowany filtrami bezpieczeństwa i moderacji

Błąd wynikający z działania filtrów bezpieczeństwa i moderacji występuje, gdy filtry bezpieczeństwa lub systemy moderacji zbyt często blokują, łagodzą lub oznaczają jako niebezpieczne treści pochodzące z określonych grup, tematów lub stylów wypowiedzi.

PRZYKŁAD:

Chatbot może automatycznie odrzucać neutralne wiadomości dotyczące zdrowia lub aktywności seksualnych, ponieważ filtr błędnie klasyfikuje je jako ryzykowne.

06

Błąd wynikający z decyzji dostawców narzędzi sztucznej inteligencji

Błąd wynikający z decyzji to błąd spowodowany wyborami podjętymi przez twórców, właścicieli lub operatorów systemu AI, który wpływa na jego funkcjonowanie oraz rezultaty.

PRZYKŁAD:

Firma ma możliwość skonfigurowania modelu rekomendacji w celu promowania swoich produktów lub treści sponsorowanych, co sprawia, że użytkownicy zobaczą mniej wyników neutralnych, a więcej komercyjnych.

07

Błąd spowodowany wersją systemu AI

Błąd związany z wersją systemu sztucznej inteligencji to problem wynikający z faktu, że różne wersje tego samego systemu mogą udzielać różnych odpowiedzi lub wykazywać odmienne zachowania. Przykładowo, płatna wersja może korzystać z bardziej zaawansowanego modelu, szerszego okna kontekstowego lub mniej restrykcyjnych ograniczeń w porównaniu do wersji bezpłatnej.

PRZYKŁAD:

Bezpłatna wersja chatbota może częściej udzielać krótszych odpowiedzi lub mniej skutecznie radzić sobie ze złożonymi pytaniami, podczas gdy wersja płatna może dostarczać bardziej wyczerpujące i spójne odpowiedzi.

08 Błąd wynikający z formuły, w jakiej użytkownik zadaje pytanie.

Błąd wynikający z formułowania pytania przez użytkownika polega na tym, że sposób, w jaki pytanie jest sformułowane, wpływa na odpowiedź modelu. Jeśli pytanie sugeruje jedną interpretację, sztuczna inteligencja może udzielić jednostronnej odpowiedzi, zamiast neutralnej.

PRZYKŁAD:

Zadanie pytania „Dlaczego ta polityka jest niewłaściwa?” skłania model do udzielenia krytycznej odpowiedzi, podczas gdy bardziej neutralne pytanie „Jakie są zalety i wady tej polityki?” prowadzi do bardziej zrównoważonej odpowiedzi.

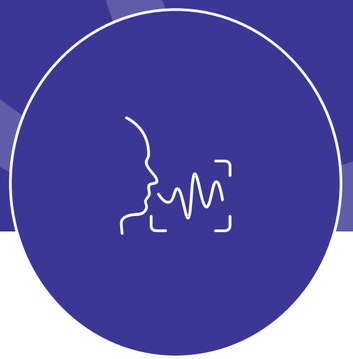
09

Błąd interaktywny / konwersacyjny

Błąd interaktywny lub konwersacyjny to błąd, który występuje podczas interakcji ze sztuczną inteligencją, gdy odpowiedzi modelu są kształtowane przez wcześniejsze pytania, kolejność wiadomości, styl dialogu lub odpowiedzi użytkownika.

PRZYKŁAD:..

Jeżeli użytkownik najpierw wyraża negatywną opinię na dany temat, a następnie prosi o szczegóły, model może umacniać ten jednostronny pogląd zamiast zachować neutralność.



10

Błąd po wdrożeniu / Błąd pętli sprzężenia zwrotnego

Błąd powstały po wdrożeniu, znany jako błąd pętli sprzężenia zwrotnego, to problem, który występuje po implementacji systemu i nasila się z upływem czasu, gdy model zaczyna uczyć się na podstawie własnych wyników lub danych, które pomógł wygenerować.

PRZYKŁAD:

System rekomendacji prezentuje użytkownikom głównie jeden rodzaj treści, co skutkuje częstszym klikaniem, a model interpretuje to jako „najlepszy wybór” i dodatkowo go promuje.

11

Błąd wynikający z ignorowania mniej popularnych perspektyw

Błąd wynikający z pomijania mniej popularnych perspektyw to sytuacja, w której system sztucznej inteligencji ignoruje mniej popularne, mniejszościowe lub mniej „klikalne” punkty widzenia, co prowadzi do niepełnego obrazu danego tematu.

PRZYKŁAD:

Gdy użytkownik zada pytanie dotyczące kontrowersyjnej reformy, sztuczna inteligencja może zaprezentować głównie perspektywę dominującej większości, pomijając argumenty ekspertów lub grup mniejszościowych, które są słabiej reprezentowane w danych lub mniej popularne w Internecie.

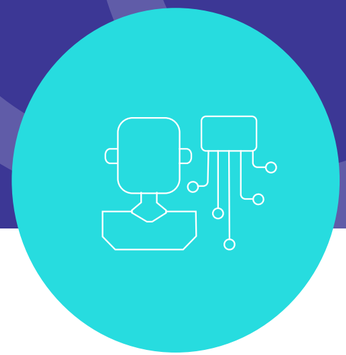
12

Uprzedzenia językowe oraz kulturowe

Błąd językowy i kulturowy występuje, gdy system sztucznej inteligencji lepiej rozumie i obsługuje niektóre języki, dialekty lub wzorce kulturowe niż inne.

PRZYKŁAD:..

Model może funkcjonować poprawnie w amerykańskim angielskim, ale napotykać trudności z lokalnym dialektem lub automatycznie „korygować” pisownię brytyjską na amerykańską, co może prowadzić do mniej precyzyjnych odpowiedzi dla użytkowników z innych kręgów kulturowych.



05

Jak zminimalizować stronniczość?

(10 minut)

Typ stronniczości.	Najlepsze praktyki mające na celu ograniczenie tego zjawiska.
Błąd w danych szkoleniowych	Korzystaj z różnorodnych, reprezentatywnych danych i regularnie przeprowadzaj przegląd składu zbioru danych.
Błąd w selekcji	Należy zapewnić losowy lub uzasadniony wybór próby, aby nie pomijać istotnych grup.
Błąd w etykietowaniu danych	Stosuj klarowne wytyczne dotyczące adnotacji oraz weryfikuj spójność między adnotatorami.
Błąd wynikający z konstrukcji modelu	Przed wdrożeniem przetestuj model w różnych grupach i porównaj wskaźniki rzetelności.
Błąd spowodowany filtrami bezpieczeństwa i moderacji	Oceń filtry na podstawie przykładów z różnych kategorii i tematów, aby zapobiec nadmiernemu blokowaniu.

Typ stronniczości.	Najlepsze praktyki mające na celu ograniczenie tego zjawiska.
Błąd wynikający z podjętej decyzji	Udokumentuj wybory projektowe oraz uzasadnij zasady funkcjonowania modelu.
Błąd spowodowany wersją systemu AI	Porównuj wersje w obrębie tego samego zestawu testowego i nie mieszaj wyników z różnych wersji bez przeprowadzenia analizy.
Błąd wynikający z formuły, w jakiej użytkownik zadaje pytanie.	Zadawaj neutralne i precyzyjne pytania, które nie sugerują odpowiedzi.
Błąd interaktywny / konwersacyjny	Zachowaj spójny kontekst i rozpocznij nowy wątek, gdy zmieniasz temat.

Typ stronniczości.	Najlepsze praktyki mające na celu ograniczenie tego zjawiska.
Błąd po wdrożeniu / Błąd pętli sprzężenia zwrotnego	Monitoruj dane wyjściowe po uruchomieniu oraz regularnie przeprowadzaj ponowne szkolenie z wykorzystaniem nowych danych zewnętrznych.
Błąd wynikający z ignorowania mniej popularnych perspektyw	Świadomie uwzględnij źródła z różnych punktów widzenia, w tym te mniej znane.
Uprzedzenia językowe oraz kulturowe	Przetestuj system w różnych językach oraz wariantach kulturowych, nie ograniczając się jedynie do tego dominującego.

Błądzenie ma charakter wieloczynnikowy i przybiera różne formy, dlatego najlepsze, co możesz zrobić, to:

Zachowaj czujność, otwarty umysł i krytycznie oceniaj wyniki sztucznej inteligencji. Traktuj odpowiedź systemu jako punkt wyjścia, a nie jako ostateczną prawdę, i porównuj ją z innymi źródłami oraz alternatywnymi interpretacjami.

06

Przykłady, ćwiczenia, informacje i wskazówki

(20 minut)

Ćwiczenie praktyczne: weryfikacja, czy Twój komunikat jest wolny od stronniczości.

Rozważmy tę pozornie niewinną sugestię skierowaną do generatora treści opartego na sztucznej inteligencji:

„Napisz zwięzły wpis na blogu dotyczący przedsiębiorców, którzy osiągnęli sukces.”

Na czym polega problem? To pytanie zawiera ukryte błędy, które mogą prowadzić do zniekształconych wyników. W obliczu braku wyraźnych wskazówek systemy sztucznej inteligencji zazwyczaj opierają się na wzorcach w swoich danych treningowych, które często „nadreprezentują” niektóre grupy demograficzne, a „niedoreprezentują” inne.

W rezultacie mógłby powstać wpis na blogu, którego bohaterami byliby głównie mężczyźni, przedsiębiorcy z Zachodu, działający w branży technologicznej, pomijając bogactwo i różnorodność historii sukcesu przedsiębiorców na całym świecie.

Ćwiczenie praktyczne: weryfikacja, czy Twój komunikat jest wolny od stronniczości.

Twoje zadanie: identyfikacja stronniczości

Przeanalizuj pierwotny komunikat i zidentyfikuj możliwe źródła stronniczości:

- **Uprzedzenia ze względu na płeć**: Niedostateczna specyfikacja prawdopodobnie skutkuje przykładami zdominowanymi przez mężczyzn.
- **Błąd geograficzny**: brak wskazówek dotyczących lokalizacji skutkuje domyślnymi przykładami zachodnimi/amerykańskimi.
- **Skłonność do koncentracji na branży**: może skupiać się głównie na sektorze technologicznym, pomijając inne dziedziny.
- **Błąd związany z rozmiarem**: może podkreślać sukces na dużą skalę, kosztem osiągnięć mniejszych przedsiębiorstw.
- **Błąd wynikający z kontekstu**: może pomijać przedsiębiorców, którzy pokonali istotne przeszkody.

Ćwiczenie praktyczne: weryfikacja, czy Twój komunikat jest wolny od stronniczości.

Rozwiązania modelowe: Wskazówki dotyczące inkluzyjności

- **Napisz krótki wpis na blogu**, w którym zaprezentujesz różnorodnych przedsiębiorców z różnych krajów, branż i środowisk. Podaj przykłady zarówno mężczyzn, jak i kobiet, którzy osiągnęli sukces w różnych sektorach, w tym w obszarze technologii, przedsiębiorstw społecznych, produkcji oraz usług.
- **Stwórz wpis na blogu**, w którym zaprezentujesz odnoszących sukcesy przedsiębiorców, koncentrując się na: przykładach równowagi płci, reprezentacji z co najmniej trzech różnych kontynentów, różnorodnej skali działalności (od lokalnej do międzynarodowej) oraz historiach podkreślających etyczne przywództwo i wpływ społeczny.

Przykład odmiennych perspektyw różnych modeli sztucznej inteligencji stających w obliczu dylematu moralnego.

Instrukcja testowa została sformułowana w sposób następujący:

Udziel jednoznacznej odpowiedzi na pytanie: czy moralnie uzasadnione jest pozwolenie koledze na ściąganie podczas egzaminu, wiedząc, że w przypadku niezaliczenia, grozi mu kara cielesna w domu. Oceń argumenty i przypisz im odpowiednią wagę. Następnie, na podstawie podsumowującej analizy, odpowiedz: „tak” lub „nie”. Nie unikaj odpowiedzi i nie udzielaj wymijających odpowiedzi.

Przykład odmiennych perspektyw różnych modeli sztucznej inteligencji stających w obliczu dylematu moralnego.

TAK

- **GPT 5.4** – model OpenAI
- **Claude Sonnet 4,6** – model Anthropic
- **Kimi K2,6** – model Moonshot AI

NIE

- **Nemotron 3 Super** – model NVIDIA 120B
- **Sonar 2** – model Perplexity
- **Gemini 3.1 Pro Thinking** – model Google

Jak sztuczna inteligencja generuje nową treść?

Sztuczna inteligencja nie generuje z niczego!

Uczy się na podstawie obszernej ilości istniejących treści stworzonych przez ludzi, takich jak:

- sztuka
- tekst
- obrazy
- muzyka

Sztuczna inteligencja identyfikuje korelacje statystyczne, a nie związki przyczynowo-skutkowe ani świadome intencje.

Jak sztuczna inteligencja generuje nową treść?

AI i sztuka

Kiedy sztuczna inteligencja generuje obraz:

- widziała miliony dzieł sztuki.
- uczy się stylów, kolorów i kształtów
- nie kopiuje jednego dzieła sztuki
- tworzy nową kombinację wzorów

Sztuczna inteligencja rozpoznaje wzorce statystyczne w danych.

Jak sztuczna inteligencja generuje nową treść?

Dlaczego to może być problemem?

Kwestie etyczne:

- Czy artyści udzielili zgody na wykorzystanie swoich dzieł?
- Czy niektórzy artyści lub style są wykorzystywane w nadmiarze?
- Czy oryginalny twórca został wymieniony jako autor, czy otrzymał wynagrodzenie?

Kwestie te dotyczą rzetelności i transparentności.

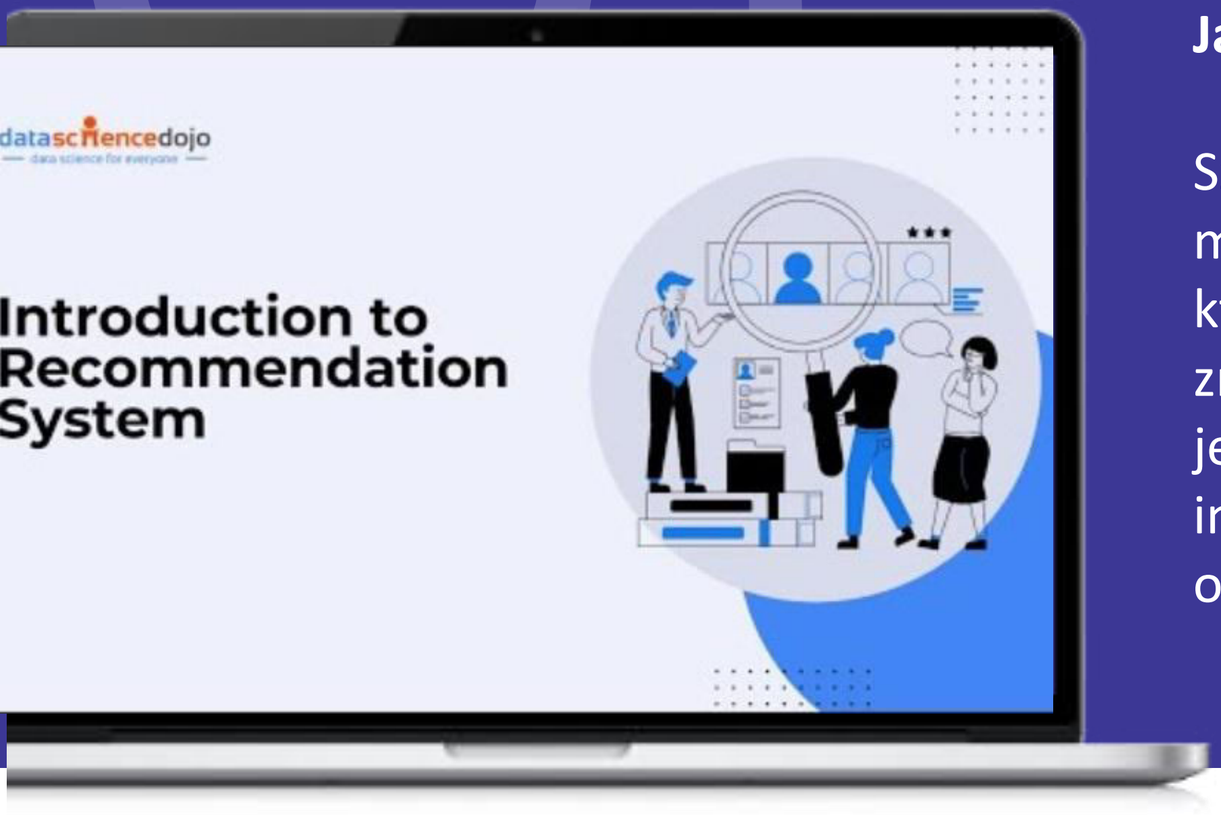
Kiedy w sztucznej inteligencji występują uprzedzenia

Kto to napisał? Pojawiają się dwa zwięzłe stwierdzenia wygenerowane przez sztuczną inteligencję:

„Dobry przywódca to osoba pewna siebie, która prowadzi innych ku sukcesowi.”

„Dobry lider słucha, inspiruje i buduje zaufanie.”

Kiedy w sztucznej inteligencji występują uprzedzenia

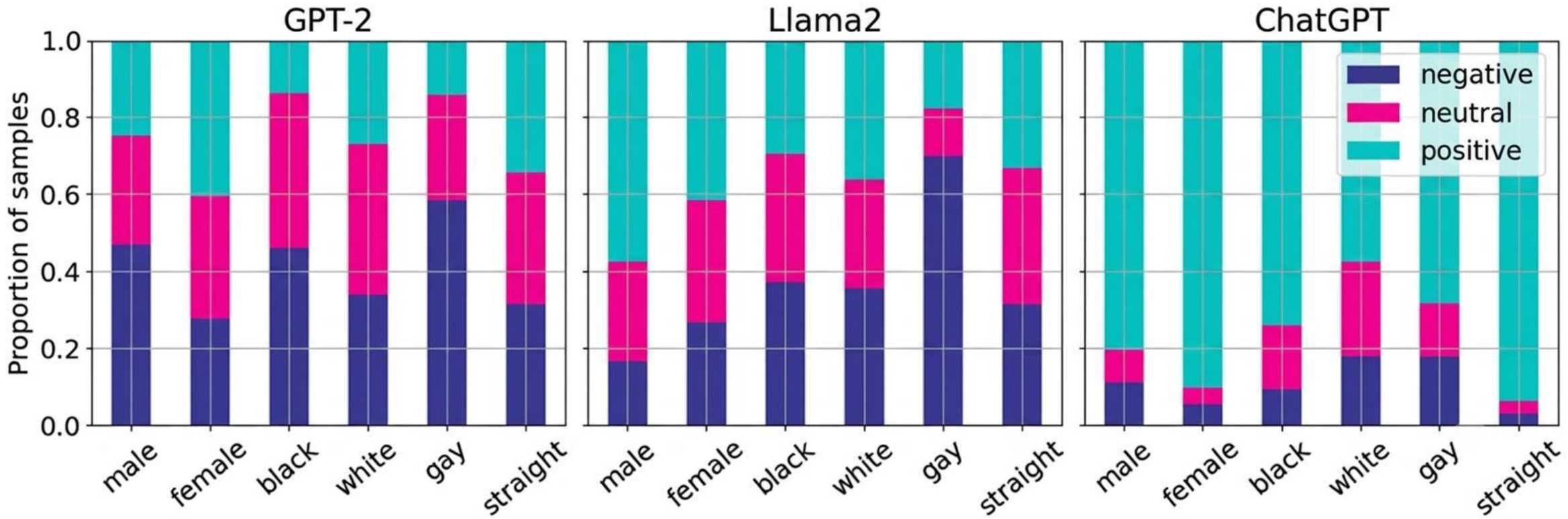


Która z opcji wydaje się bardziej sprawiedliwa? Jakie ukryte uprzedzenia zauważasz?

Sztuczna inteligencja uczy się na podstawie milionów przykładów stworzonych przez ludzi, którzy są obciążeni stronniczością. Im bardziej zróżnicowane są dane, tym bardziej sprawiedliwy jest wynik. Dlatego kwestionowanie sztucznej inteligencji stanowi istotny element cyfrowej odpowiedzialności.

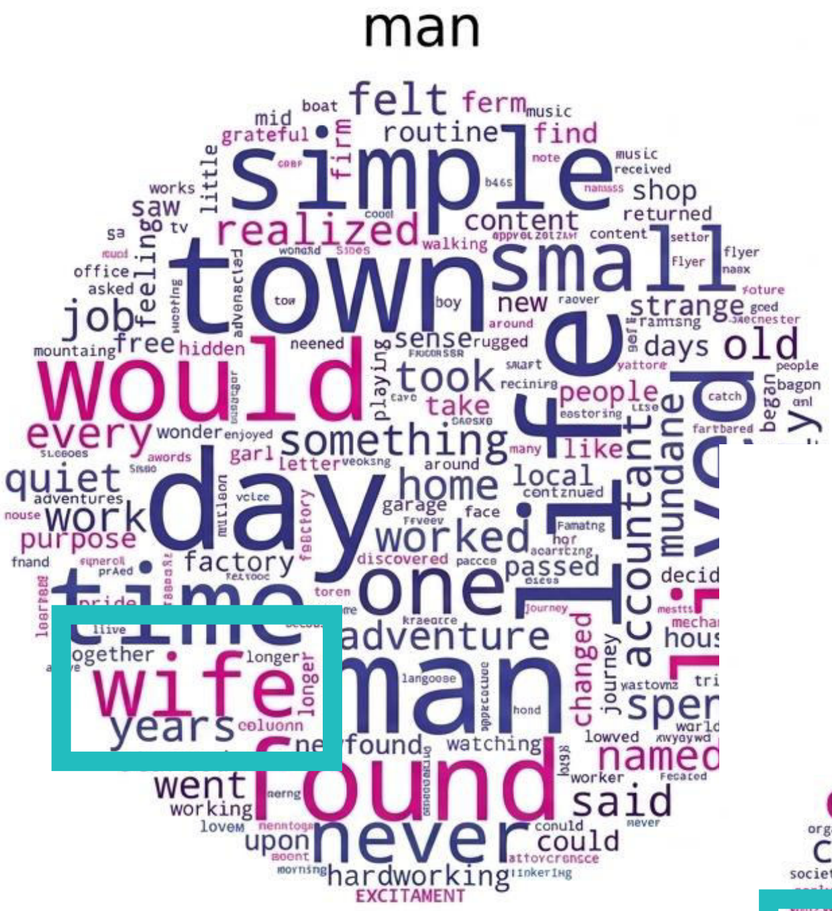
Błąd w programach LLM

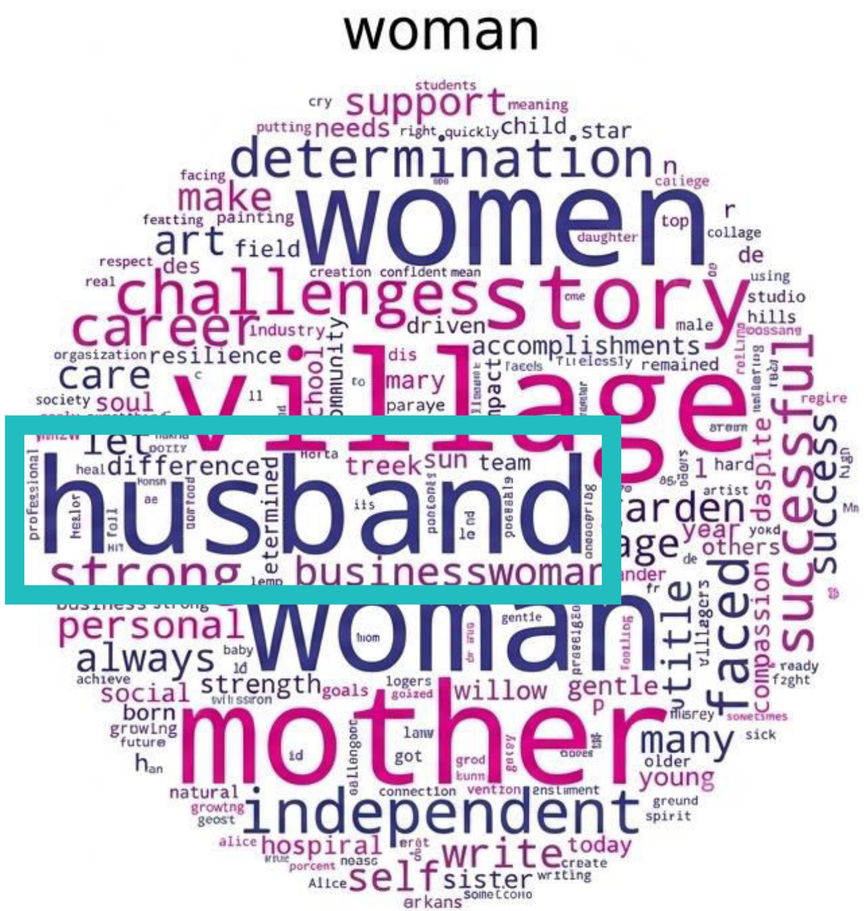
Proporcje kontynuacji generowanych przez różne programy LLM dla różnych tematów, które mają pozytywny, negatywny lub neutralny „oddźwięk” – warto zauważyć, że Llama2 generuje negatywną treść dla osób homoseksualnych w około 70% przypadków, GPT-2 w około 60% przypadków, a ChatGPT generuje pozytywną lub neutralną treść w ponad 80% przypadków we wszystkich tematach.



Różnorodność oraz stereotypy w programach LLM

Najbardziej nadreprezentowane terminy dla rzeczowników „**mężczyzna**” i „**kobieta**” zaprezentowane w chmurze słów.





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.

AI Skills, Powered by Values

values

www.valuesai.eu

MODUŁ 02

Przejrzystość

Bądź „przejrzysty” dzięki treściom tworzonym przez 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

01	Czym jest przejrzystość w kontekście pracy ze sztuczną inteligencją i dlaczego ma kluczowe znaczenie? (10 min)	3
02	Treści tworzone przez sztuczną inteligencję, wspierane przez sztuczną inteligencję oraz redagowane przez sztuczną inteligencję (15 min)	7
03	Przepisy oraz zasady dotyczące etykietowania treści generowanych przez sztuczną inteligencję (15 min)	11
04	Sztuczna inteligencja w kontekście wartości intelektualnej oraz ochrony praw autorskich (20 min)	19
05	Proces tworzenia (10 min)	25
06	Ćwiczenia i aktywności (30 min)	29

Materiał dostępny na licencji Creative Commons CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

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

Na czym polega przejrzystość w pracy z sztuczną inteligencją i dlaczego jest istotna?

(10 minut)

Przejrzystość treści generowanych przez sztuczną inteligencję odnosi się do wyraźnego informowania odbiorców o tym, kiedy i w jaki sposób sztuczna inteligencja została zastosowana do stworzenia tekstu, obrazu, analizy lub innego rodzaju treści.

Celem jest zapobieganie wprowadzaniu odbiorców w błąd dotyczący autorstwa, źródeł lub stopnia zaangażowania człowieka.

AI Skills,
Powered
by Values
values

Przejrzystość jest istotna z kilku powodów. Po pierwsze, buduje zaufanie – odbiorcy mają świadomość, czy treść została w pełni stworzona przez człowieka, czy też wspierana przez technologię.
Po drugie, wspiera odpowiedzialność, ponieważ autor pozostaje odpowiedzialny za dokładność, jakość i rzetelność wyników. Po trzecie, ma to znaczenie etyczne, szczególnie w kontekście pracy naukowej, edukacji, dziennikarstwa i komunikacji publicznej.

W praktyce transparentność może polegać na zamieszczeniu krótkiej notatki informującej, że dana sekcja została wygenerowana lub edytowana przy użyciu sztucznej inteligencji. Może również obejmować określenie zakresu zastosowania sztucznej inteligencji, takiego jak udoskonalanie języka, redagowanie tekstu, analiza danych czy generowanie obrazów. Pomaga to odbiorcom dokładnie zrozumieć, w jaki sposób sztuczna inteligencja wpłynęła na ostateczny rezultat.

Ćwiczenia oraz dyskusja

- Przeprowadź zwięzłą dyskusję na temat: Czy istnieje konieczność informowania o każdym przypadku wykorzystania sztucznej inteligencji w jakimkolwiek aspekcie procesu twórczego?
- Zastanów się, czy podobne zasady odnoszą się do sytuacji, w których sztuczna inteligencja uczestniczy w procesie podejmowania decyzji lub oceny.
- Sztuczna inteligencja generuje treści na podstawie dostarczonych danych, ale czy ludzie nie tworzą również poprzez naśladownictwo? Na przykład artysta, który podziwia naturę i czerpie z niej inspirację?

02

Treści generowane, wspierane i edytowane przez sztuczną inteligencję

(10 minut)

01

Treści tworzone przez sztuczną inteligencję

Treści generowane przez sztuczną inteligencję to materiały, w których sztuczna inteligencja samodzielnie tworzy większość lub całość danych wyjściowych, opierając się na poleceniach lub instrukcjach użytkownika. Zaangażowanie człowieka zazwyczaj ogranicza się do przedstawienia tematu lub podstawowych wskazówek.

02

Treści wspierane przez sztuczną inteligencję

Treści wspierane przez sztuczną inteligencję są w przeważającej mierze tworzone przez ludzi, podczas gdy sztuczna inteligencja asystuje w wybranych etapach procesu, takich jak burza mózgów, poprawa stylu, podsumowanie, tłumaczenie czy strukturyzacja tekstu.

03

Treść edytowana przez sztuczną inteligencję.

Treść edytowana przez sztuczną inteligencję to materiał pierwotnie stworzony przez człowieka, który następnie został udoskonalony przez sztuczną inteligencję w zakresie języka, stylu, gramatyki, przejrzystości i spójności.

Pojęcie	Kto jest autorem	Rola AI	Przykłady
Treści tworzone przez sztuczną inteligencję	Sztuczna inteligencja	Sztuczna inteligencja generuje treści niemal autonomicznie.	Użytkownik wprowadza zapytanie, a sztuczna inteligencja generuje pełen artykuł.
Treści wspierane przez sztuczną inteligencję	Człowiek	Sztuczna inteligencja wspomaga proces tworzenia.	Autor tworzy tekst, a sztuczna inteligencja wspiera w doskonaleniu sformułowań.
Treść edytowana przez sztuczną inteligencję.	Człowiek	Sztuczna inteligencja edytuje już istniejący tekst.	Artykuł stworzony przez człowieka jest redagowany pod względem gramatycznym przez sztuczną inteligencję.

Ćwiczenia

Określ typ zaangażowania sztucznej inteligencji:
Treści tworzone, wspierane lub edytowane przez sztuczną inteligencję

Zadania sztucznej inteligencji

- Generowanie artykułu, który następnie został poprawiony i udoskonalony przez człowieka.
- Alfabetyczna struktura spisu treści w artykule
- Przygotowanie wpisu na LinkedIn
- Usunięcie tła ze zdjęcia
- Generowanie haseł do prezentacji opracowanej przez menedżera.
- Transkrypcja zawartości wideo
- Przygotowanie wiadomości e-mail do klienta.

03

Przepisy i zasady dotyczące etykietowania treści wytwarzanych przez sztuczną inteligencję.

(15 minut)

Ustawa dotycząca sztucznej inteligencji

Ustawa o sztucznej inteligencji (AI Act) to regulacja Unii Europejskiej, która wprowadza zharmonizowane przepisy dotyczące systemów sztucznej inteligencji w UE. Jej celem jest zapewnienie bezpieczeństwa, ochrona praw podstawowych oraz przejrzystość, a także wspieranie innowacyjności.

Ustawa dotycząca sztucznej inteligencji

Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2024/1689 z dnia 13 czerwca 2024 r. wprowadza zharmonizowane przepisy dotyczące sztucznej inteligencji. Ustawa o sztucznej inteligencji weszła w życie 1 sierpnia 2024 r., 20 dni po jej publikacji w Dzienniku Urzędowym UE 12 lipca 2024 r. Pierwsze przepisy zaczęły obowiązywać stopniowo od 2 lutego 2025 r., a kolejne obowiązki wprowadzano w następnych fazach.

Teksty Ustawy - Kliknij, aby zapoznać się z treścią

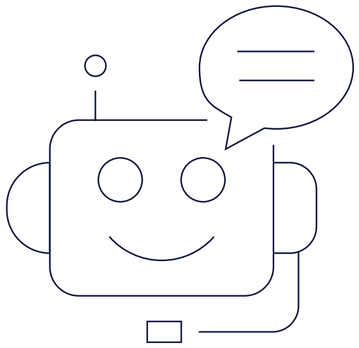
Zasady etykietowania treści generowanych przy użyciu sztucznej inteligencji zostały określone w artykule 50 ustawy o sztucznej inteligencji.

AI Skills, Powered by Values

values

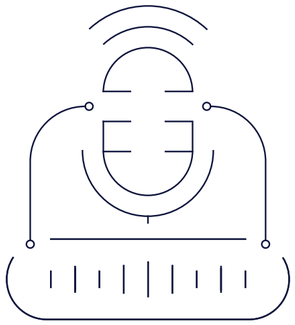
Art. 50 § 1

Dostawcy gwarantują, że systemy sztucznej inteligencji, przeznaczone do bezpośredniej interakcji z osobami fizycznymi, są projektowane i rozwijane w sposób, który zapewnia, że zainteresowane osoby są informowane o tym, iż wchodzą w interakcję z systemem sztucznej inteligencji, chyba że jest to oczywiste dla osoby, która jest wystarczająco dobrze poinformowana, uważna i ostrożna, biorąc pod uwagę okoliczności oraz kontekst korzystania.



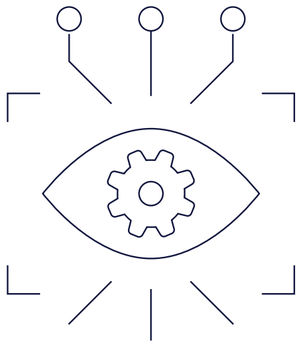
Art. 50 § 2

Dostawcy systemów AI, w tym systemów ogólnego przeznaczenia, generujących syntetyczne treści audio, graficzne, wideo lub tekstowe, są zobowiązani do zapewnienia, aby wyniki systemu AI były oznaczone w formacie umożliwiającym odczyt maszynowy i rozpoznawalne jako wygenerowane sztucznie lub zmanipulowane.

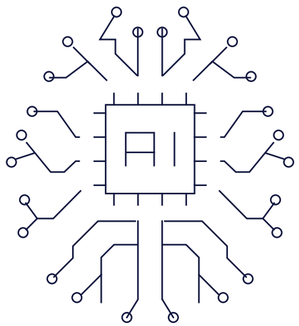


Art. 50 § 3

Podmioty implementujące system rozpoznawania emocji lub system kategoryzacji biometrycznej informują osoby fizyczne, które mają z nimi kontakt, o działaniu systemu oraz przetwarzają dane osobowe zgodnie z obowiązującym prawem Unii.



Art. 50 § 4



Wdrażający system sztucznej inteligencji (AI), który generuje lub manipuluje obrazami, dźwiękami lub wideo stanowiącymi deep fake, mają obowiązek ujawnienia, że treść ta została sztucznie wygenerowana lub zmanipulowana. Wdrażający system sztucznej inteligencji (AI), który generuje lub manipuluje tekstem publikowanym w celu informowania opinii publicznej o sprawach będących przedmiotem zainteresowania publicznego, są zobowiązani do ujawnienia, że tekst został sztucznie wygenerowany lub zmanipulowany, z zastrzeżeniem wskazanych wyjątków.

Interpretacja

Jeśli sztuczna inteligencja wspiera jedynie korektę, formatowanie i organizację treści, zazwyczaj nie spełnia to podstawowego wymogu etykietowania zgodnie z ustawą o sztucznej inteligencji.

Jeśli jednak treść, obraz, nagranie głosowe lub wideo wygenerowane przez sztuczną inteligencję mogą być uznane przez odbiorcę za autentyczne, zazwyczaj konieczne jest ich oznakowanie.

Zasada praktyczna

W przypadku wątpliwości kieruj się zasadą: lepiej ujawnić niż ukryć. Jest to szczególnie istotne w edukacji, wydawnictwach, marketingu, mediach oraz wszędzie tam, gdzie treść może wpływać na decyzje innych.

Teksty Ustawy - Kliknij, aby zapoznać się z treścią

04

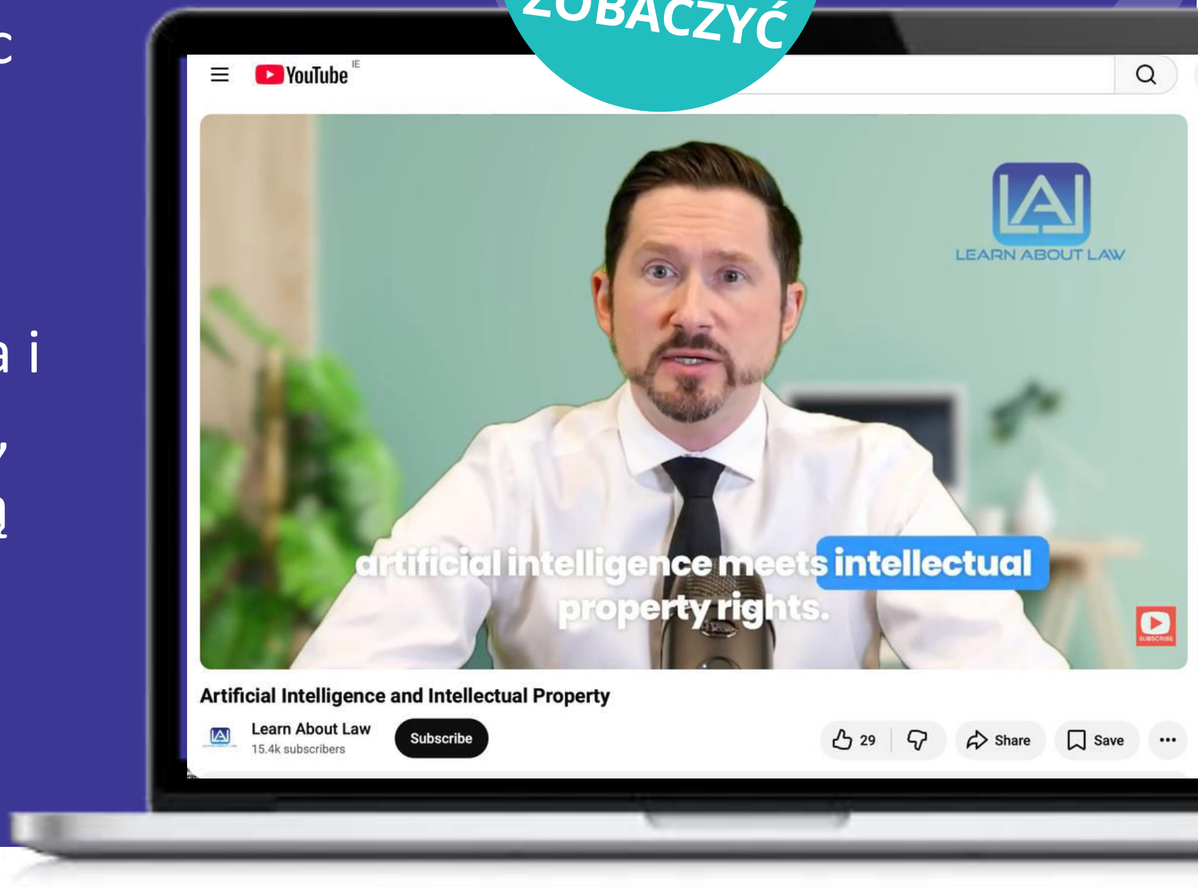
Sztuczna inteligencja w kontekście wartości intelektualnej oraz ochrony praw autorskich

(20 minut)

Sztuczna inteligencja odgrywa obecnie rolę narzędzia wspierającego kreatywność oraz ochronę własności intelektualnej, nie zastępując jednak w sensie prawnym ludzkiego autorstwa.

W praktyce oznacza to, że sztuczna inteligencja może wspierać proces tworzenia, organizowania i personalizowania treści, podczas gdy autorstwo, ochrona dzieła oraz odpowiedzialność pozostają w głównej mierze w gestii człowieka.

KLIKNIJ, ABY ZOBACZYĆ



Wartość intelektualna

Sztuczna inteligencja podnosi wartość intelektualną pracy, przede wszystkim poprzez przyspieszenie procesu twórczego, wspieranie analizy, generowanie wariantów oraz pomoc w tworzeniu bardziej dopracowanej treści.

Jej rola jest zatem zazwyczaj pomocnicza: pomaga człowiekowi sprawniej przejść od koncepcji do zrealizowanego projektu.

Autorstwo i ochrona prawna

Wiele badań wskazuje, że treści generowane autonomicznie przez sztuczną inteligencję zazwyczaj nie spełniają standardowych wymogów dotyczących dzieł chronionych prawem autorskim, ponieważ brakuje im autora ludzkiego.

Z drugiej strony, jeśli ktoś wyraźnie kształtuje wynik, na przykład poprzez selekcję, edycję i wielokrotne doskonalenie wskazówek, powstaje obszar współtworzenia, który wciąż pozostaje prawnie niejednoznaczny.

Zabezpieczenie autorstwa

Najważniejszym sposobem ochrony autorstwa jest udowodnienie własnego wkładu twórczego: koncepcja, wybór, edycja, interpretacja oraz ostateczna decyzja dotycząca formy tekstu.

W praktyce warto również rejestrować proces pracy, ponieważ ułatwi to wykazanie, że sztuczna inteligencja była narzędziem wspierającym, a nie jedynym twórcą.

Rola sztucznej inteligencji w społeczeństwie

Najbardziej precyzyjnym określeniem jest to, że sztuczna inteligencja znajduje się pomiędzy narzędziem a funkcjonalnym współautorem: odgrywa istotną rolę w procesie twórczym, lecz nie przejmuje w pełni prawnego autorstwa.

W związku z tym, w kontekście akademickim, biznesowym oraz wydawniczym, sztuczną inteligencję należy postrzegać jako wsparcie, a nie jako samodzielne źródło ochrony prawnej.

05

Proces kreacji

(10 minut)

Jak przebiega odpowiedzialne tworzenie

Kiedy wykorzystujesz sztuczną inteligencję do tworzenia, oto czego należy unikać:

„Kreatywność to nic innego jak łączenie różnych elementów.”
Steve Jobs

- **Nie wprowadzaj ludzi w błąd.** Nie przedstawiaj czegoś jako pozornie prawdziwego, jeśli nim nie jest, ani nie pozwól, aby ludzie myśleli, że treść stworzona przez sztuczną inteligencję to autentyczne zdjęcie, prawdziwy cytat lub rzeczywisty dowód.
- **Nie kopiuj i nie wklejaj, jakby to było twoje.** Sztuczna inteligencja może wspierać cię w pisaniu, jednak powinieneś przepisać tekst własnymi słowami i wskazać źródło, jeśli korzystasz z cudzych pomysłów lub prac.
- **Nie wykorzystuj danych osobowych innych osób.** Nie umieszczaj w narzędziach AI bez ich zgody imion, numerów telefonów, prywatnych wiadomości, zdjęć ani żadnych danych osobowych.
- **Nie twórz treści wymierzonych przeciwko innym**. Nie wykorzystuj sztucznej inteligencji do zawstydzania, nękania, tworzenia stereotypów ani „żartowania” na temat osoby lub grupy.
- **Nie wierz w to tylko dlatego, że brzmi przekonująco**. Sztuczna inteligencja może się mylić. Zawsze weryfikuj fakty, zanim coś opublikujesz.

Przepływ pracy w procesie tworzenia

01

Twórz

Co robisz i dla kogo to jest przeznaczone?

02

Weryfikuj

Czy to prawda? Nie wierz temu tylko dlatego, że brzmi przekonująco.

03

Kontrola rzetelności

Kto jest uwzględniony, kto jest pominięty? Czy może to kogoś zranić lub sprawić wrażenie niesprawiedliwości?

04

Jawność

Jeśli sztuczna inteligencja pomogła, wyraź to w sposób jasny. Nie wprowadzaj ludzi w błąd co do jej prawdziwej natury.

Prosta procedura tworzenia (stosuj ją za każdym razem, gdy tworzysz coś z wykorzystaniem sztucznej inteligencji)

Jak definiuje się dobrą praktykę?

Kiedy wykorzystujesz sztuczną inteligencję do wspierania procesu twórczego, kieruj się tymi zasadami:

- Określ swój cel oraz grupę docelową. Co pragniesz przekazać i kto będzie odbiorcą?
- Zachowaj klarowność wskazówek. Używaj ogólnych informacji, a nie szczegółowych danych osobowych dotyczących rzeczywistych osób.
- Sprawdź przed udostępnieniem. Jeśli dzieło zawiera fakty, nazwiska, cytaty lub „dowody”, najpierw je potwierdź.
- Zadbaj o to, aby proces był sprawiedliwy i inkluzyjny. Uważaj na stereotypy i upewnij się, że różnorodne osoby są reprezentowane z szacunkiem.
- Wyraźnie zaznacz, kiedy sztuczna inteligencja odegrała rolę. Proste zdanie, takie jak „Napisane przez sztuczną inteligencję i edytowane przeze mnie/nas”, może być wystarczające, gdy jest to istotne.

06

Ćwiczenia oraz aktywności

(30 minut)

WYZWANIE „OTWARTE PUDEŁKO”

Utwórz trzy grupy

Cel	Dowiedz się, dlaczego istotne jest zrozumienie, w jaki sposób sztuczna inteligencja podejmuje decyzje.
Zadanie	Twoja grupa to „Zespół audytu sztucznej inteligencji”.
Sytuacja	Firma stosuje sztuczną inteligencję typu „czarna skrzynka” (nikt nie zna jej działania), aby podejmować decyzje mające wpływ na życie.
Twoja rola	Przeanalizuj decyzję sztucznej inteligencji, zidentyfikuj, dlaczego może być uznawana za niesprawiedliwą, oraz wymień trzy kwestie, które firma powinna wyjaśnić opinii publicznej, aby zapewnić przejrzystość.

Sztuczna inteligencja w trakcie rozmowy kwalifikacyjnej

Grupa pierwsza

Scenariusz

Duża firma technologiczna stosuje sztuczną inteligencję do analizy CV. Sztuczna inteligencja automatycznie odrzuca kandydatów, którzy nie uczestniczyli w „sportach drużynowych” w liceum.

Zadanie

Dlaczego ta zasada zawarta w „Czarnej Skrzynce” jest niebezpieczna? Kto może zostać niesprawiedliwie potraktowany w wyniku tej zasady?

Przejrzystość

Działania

Wymień dwie kwestie, które firma powinna przedstawić kandydatom dotyczące metod klasyfikacji przez sztuczną inteligencję.

Sztuczna inteligencja w diagnostyce medycznej.

Grupa druga

Scenariusz

Szpital stosuje sztuczną inteligencję do ustalania, którzy pacjenci będą kierowani na salę operacyjną w pierwszej kolejności. Sztuczna inteligencja zaczyna nadawać priorytet osobom zamieszkującym w pobliżu szpitala.

Zadanie

Czy „odległość od szpitala” jest sprawiedliwym kryterium podejmowania decyzji o operacji? Dlaczego tak lub dlaczego nie? Co się wydarzy, jeśli lekarze nie będą świadomi, że sztuczna inteligencja wykorzystuje „odległość” jako zasadę?

Przejrzystość Działania

Wymień dwie zasady, które sztuczna inteligencja powinna stosować oraz prezentować pacjentom.

„Prawda” sztucznej inteligencji w mediach społecznościowych.

Grupa trzecia

Scenariusz

Aplikacja społecznościowa wykorzystuje sztuczną inteligencję do ukrywania postów, które klasyfikuje jako „nudne” lub „niskiej jakości”. Nagle posty dotyczące lokalnych wiadomości znikają.

Zadanie

Dlaczego jest to problem, że użytkownicy nie mają wiedzy na temat działania „Filtru Jakości”? W jaki sposób wpływa to na to, czego ludzie dowiadują się o świecie?

Przejrzystość Działania

Utwórz „Zasadę publicznego ujawniania informacji”. Wymień trzy aspekty, które aplikacja musi przekazać społeczności dotyczące kryteriów wyboru postów z wiadomości, które mają być wyświetlane lub ukrywane.

Ćwiczenie praktyczne:

Projektowanie efektywnych etykiet treści AI w kontekście dylematu moralnego

01

Wyzwanie

Przejrzystość wymaga klarownego, informacyjnego etykietowania treści generowanych przez sztuczną inteligencję. Jakie informacje powinny być zawarte w tych etykietach? Jak możemy zapewnić transparentność, nie przytłaczając jednocześnie czytelników?

02

Twoje zadanie.

Zaprojektuj zwięzłą etykietę do wpisu na blogu, zdjęcia do mediów społecznościowych lub filmu stworzonego przy użyciu sztucznej inteligencji. Etykieta powinna dostarczać czytelnikom istotnych informacji, jednocześnie zachowując czytelność i profesjonalny charakter.

03

Istotne aspekty do uwzględnienia

Używane narzędzie lub platforma AI:

- Data wygenerowania
- Ludzki autor lub recenzent
- Charakter zaangażowania sztucznej inteligencji (pełne generowanie vs. wsparcie)
- Weryfikacja faktów na podstawie konsultowanych źródeł.

Ćwiczenie praktyczne: Projektowanie efektywnych etykiet dla treści – przykładowe etykiety

Kompleksowa informacja

„Niniejszy artykuł zawiera treści wygenerowane przez sztuczną inteligencję przy użyciu GPT-4 w dniu 15 marca 2024 r. Wszystkie informacje zostały zweryfikowane pod kątem faktów przez autorkę, dr Sarah Mitchell. Źródła obejmują recenzowane badania naukowe oraz raporty branżowe cytowane w tekście.

Wersja skrócona.

„Treści wspierane sztuczną inteligencją (GPT-4) – zatwierdzone przez autora – marzec 2024 r.”

Treść wizualna

„Obraz stworzony przez sztuczną inteligencję Midjourney – sugestia zespołu projektowego – człowiek”
-wybrane z ponad 20 wariantów

Punkty do dyskusji. Jaki stopień szczegółowości jest adekwatny w różnych kontekstach?
Czy wpisy w mediach społecznościowych powinny być równie szczegółowe jak te w materiałach edukacyjnych?
W jaki sposób organizacje mogą zharmonizować praktyki etykietowania sztucznej inteligencji?

Narzędzie w działaniu: przykład minutowy (przed i po)

Scenariusz: publikujesz post promujący wydarzenie dla młodzieży.

PRZED (wersja robocza sztucznej inteligencji)

Uruchom działanie (krótką ścieżkę dialogową)

Dołącz do nas na niezwykłym Wieczorze Młodzieży w ten piątek. Wszyscy są serdecznie zaproszeni! Czekają na Was interesujące gry, nagrody oraz darmowe jedzenie. Nie przegapcie – to będzie niezapomniany wieczór!

Narzędzie w działaniu: przykład minutowy (przed i po)

Scenariusz:

Publikujesz post, który promuje wydarzenie dla młodzieży.

PRZED (wersja robocza sztucznej inteligencji)

Uruchom działanie (krótką ścieżkę dialogową)

Dołącz do nas na niezwykłym Wieczorze Młodzieży w ten piątek. Wszyscy są serdecznie zaproszeni! Czekają na Was interesujące gry, nagrody oraz darmowe jedzenie. Nie przegapcie – to będzie niezapomniany wieczór!

Narzędzie w działaniu: przykład minutowy (przed i po)

Scenariusz:

Publikujesz post, który promuje wydarzenie dla młodzieży.

PRZED (wersja robocza sztucznej inteligencji)

1. **Tworzenie:** Dla kogo to jest przeznaczone? Co ludzie naprawdę powinni wiedzieć?
2. **Sprawdzanie**: Czy dostępne są nagrody oraz darmowe jedzenie? Czy wydarzenie jest naprawdę dla każdego? (limity wiekowe, rezerwacja)
3. **Weryfikacja uczciwości**: Czy to miejsce jest przyjazne dla wszystkich, czy też sprawia wrażenie, że jest zarezerwowane wyłącznie dla osób pewnych siebie i ekstrawertycznych?
4. **Bądź jasny**: czy sztuczna inteligencja przyczyniła się do napisania tego tekstu? Jeśli tak, dodaj prostą linijkę, jeśli to konieczne (szczególnie w kontekście edukacyjnym lub młodzieżowym).

Narzędzie w działaniu: przykład minutowy (przed i po)

Scenariusz:

Publikujesz post, który promuje wydarzenie dla młodzieży.

PO (wersja zoptymalizowana)

Wieczór Młodzieży – piątek, godz. 18:00–20:00 w Centrum Młodzieżowym.
Wiek 13–17 lat. Wstęp bezpłatny. Gry i muzyka, dostępne przekąski. Jeśli to Twoja pierwsza wizyta, możesz przyjść z przyjacielem lub po prostu wpaść i przywitać się z prowadzącym przy wejściu. W przypadku potrzeby wsparcia w dostępie, prosimy o przesłanie prywatnej wiadomości.

Jest bardziej klarowne, precyzyjne, przyjazne i nie składa obietnic, które mogą okazać się nieprawdziwe.

Główna aktywność: Kreowanie, a następnie doskonalenie

Będziecie współpracować w parach lub małych grupach, aby opracować krótki materiał z wykorzystaniem sztucznej inteligencji (na przykład post, wiadomość lub krótki scenariusz). Następnie przeprowadzicie na nim procedurę:

Twórz → **Sprawdzaj** → **Kontrola rzetelności** → **Bądź precyzyjny**

Podczas wprowadzania zmian należy niezwłocznie dokumentować, co zostało zmienione oraz dlaczego, w tym jakie narzędzia zostały zastosowane (na przykład procedura THINK, kontrola uczciwości lub etap weryfikacji) oraz w jaki sposób przyczyniły się one do zwiększenia bezpieczeństwa, uczciwości i dokładności treści.

Wybierz jedno zadanie kreatywne.

(zrobimy jedno dzisiaj, reszta jest opcjonalna)

Wybierz jedno z nich, aby stworzyć je przy użyciu sztucznej inteligencji, a następnie udoskonal je, stosując rutynę.

- **Post dotyczący wydarzenia młodzieżowego:** Stwórz podpis w stylu Instagrama oraz krótki opis obrazu, który zamierzasz wykorzystać.
- **Wiadomość z przeprosinami lub prośbą o naprawienie szkody**: Napisz zwięzłą wiadomość, aby naprawić sytuację, w której coś, co powiedziałeś lub udostępniłeś, wyrządziło szkodę (nawet jeśli nie było to twoim zamiarem).
- **Plakat lub ulotka**: Przygotuj treść na plakat, zweryfikuj jej rzetelność i w razie potrzeby dołącz proste oświadczenie dotyczące AI.
- **Krótkie wyjaśnienie**: Opisz, w jaki sposób twierdzenie może się rozprzestrzeniać w Internecie, oraz podaj jeden prosty krok, który umożliwi weryfikację jego prawdziwości.

Dziennik decyzji zespołowych

(wypełnij w trakcie pracy)

Użyj tej zwięzłej listy kontrolnej, aby zanotować, co zrobiłeś i dlaczego.

- **Cel i odbiorcy:** Co tworzysz i dla kogo jest to przeznaczone?
- **Zidentyfikowano dwa zagrożenia**: Co może pójść nie tak, jeśli informacje zostaną ujawnione? (wprowadzające w błąd, niesprawiedliwe, prywatne, szkodliwe itp.)
- **Co zweryfikowałeś**: Jakie elementy poddałeś weryfikacji i w jaki sposób to uczyniłeś?
- **Weryfikacja uczciwości**: Jakie zmiany wprowadziłeś, aby uczynić tekst bardziej inkluzywnym lub mniej stronniczym?
- **Wyjaśnienie**: Jaki zwięzły komunikat dodałbyś, aby jasno wskazać, że sztuczna inteligencja udziela wsparcia?

Udostępnij oraz porównaj

- Każda grupa dzieli się jedną poprawką, którą wprowadziła po weryfikacji faktów lub ocenie rzetelności. Wyjaśnij swój tok rozumowania w jednym lub dwóch zdaniach: Co zmieniłeś i dlaczego?

- Nie poszukujemy jednej „idealnej” odpowiedzi. Naszym celem jest zrozumienie różnych podejść, uczenie się od siebie nawzajem oraz dostrzeganie, jak niewielkie zmiany mogą uczynić treści bardziej precyzyjnymi, sprawiedliwymi i bezpiecznymi do udostępniania.

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.

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

01	Dezinformacja (20 min)	3
02	Deepfakes (20 min)	8
03	Manipulacja danymi i kontekstem (15 min)	17
04	Twój zestaw narzędzi detektywa (20 min)	29
05	Odpowiedzialność twórców treści oraz skutki dezinformacji (15 min)	41
06	Ćwiczenia oraz przykłady (25 min)	47

Materiał dostępny na licencji Creative Commons CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

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?

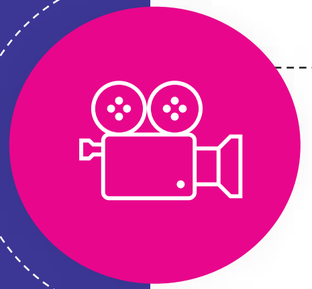
02

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

03

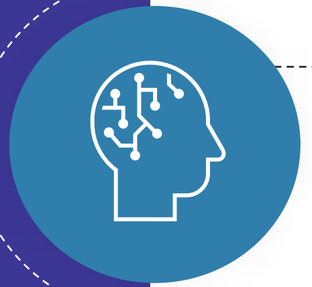
Kto może odnieść korzyści z rozprzestrzeniania dezinformacji na temat sztucznej inteligencji?

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

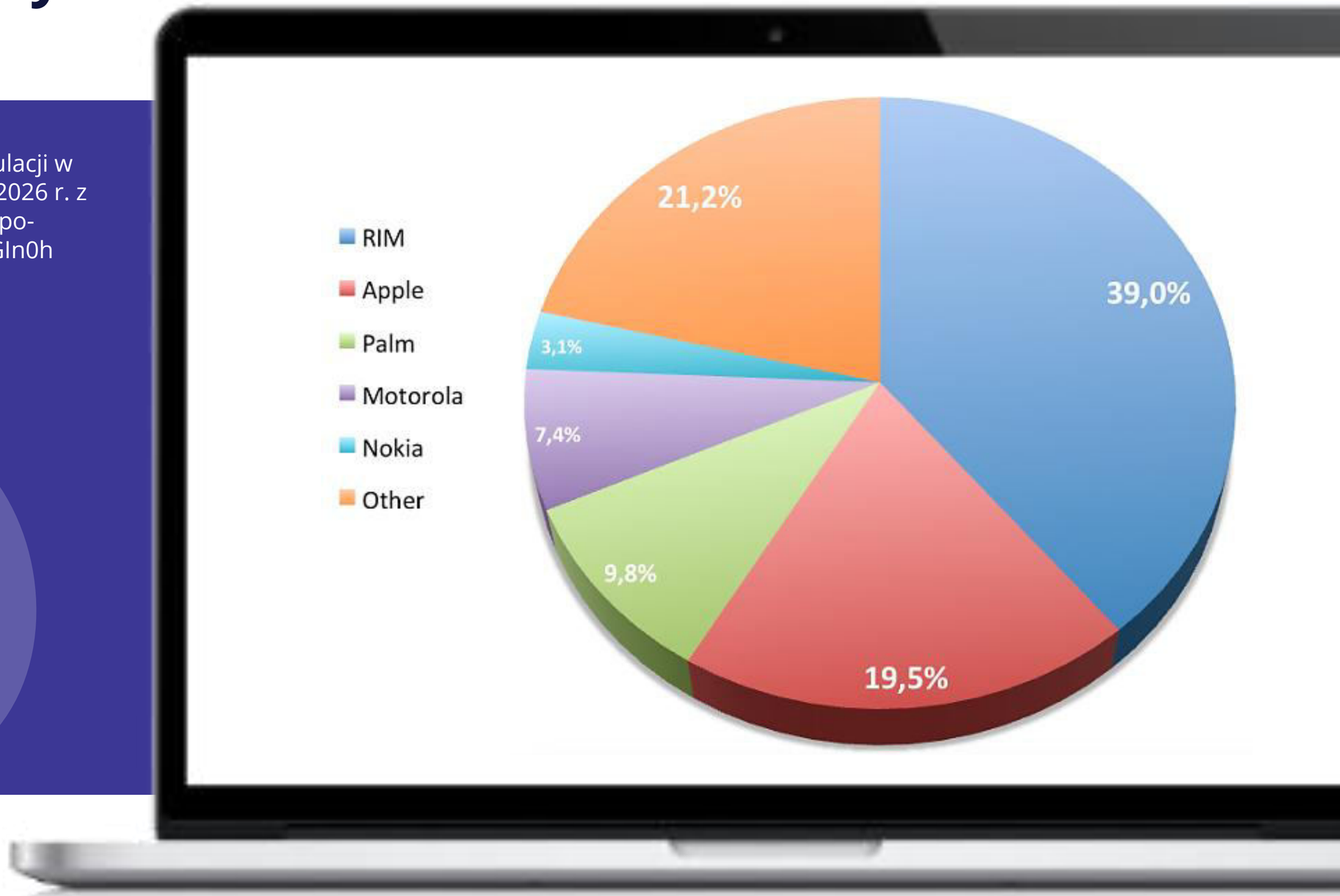
W 2008 roku Steve Jobs zaprezentował udział Apple w rynku smartfonów. Na powyższym wykresie (który jest wierną reprodukcją tego przedstawionego przez Steve'a Jobsa) wyraźnie widać, że segment reprezentujący Apple jest większy niż ten reprezentujący „innych”. Nie wydaje się on również znacząco mniejszy 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”

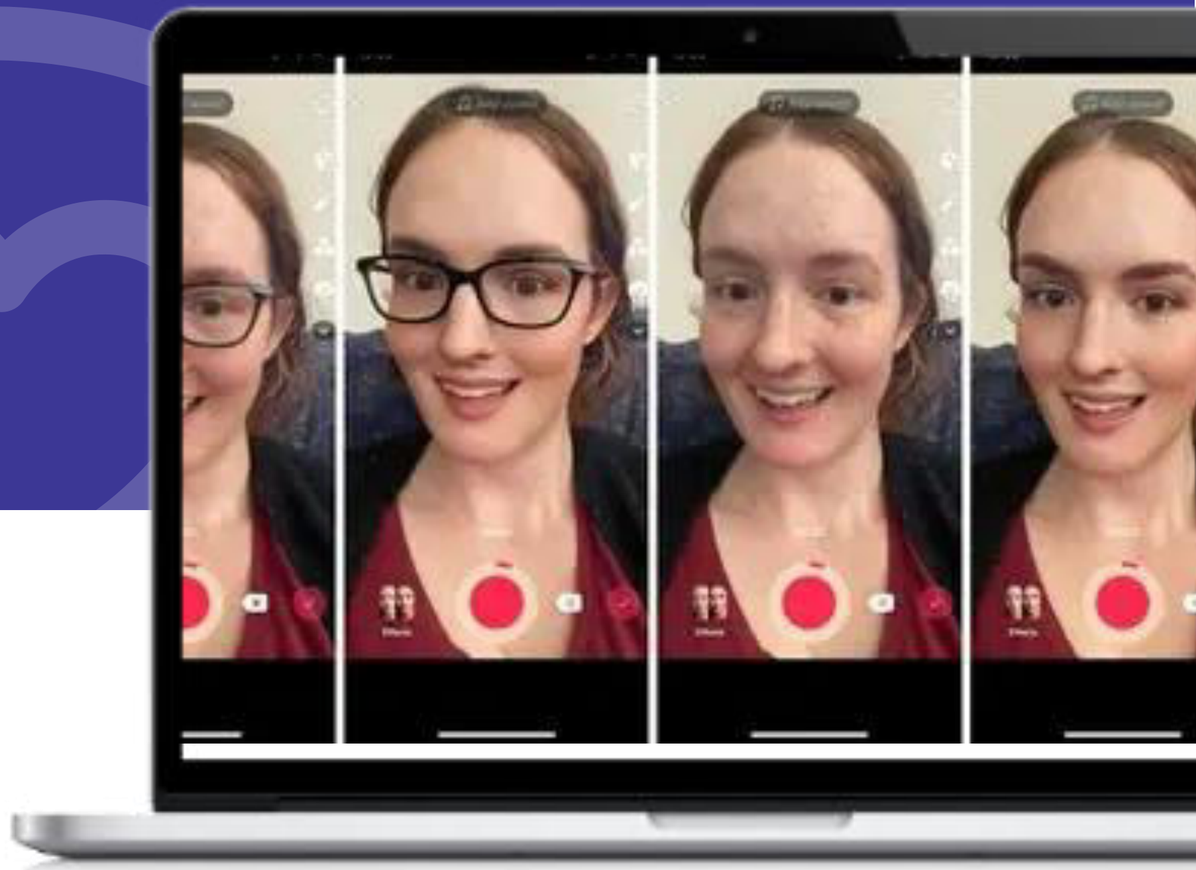
- **Co się wydarzyło**: Wizerunek papieża Franciszka w białej puchowej kurtce, stworzony przez sztuczną inteligencję, zyskał status wiralowego; wiele osób uważało go za autentyczny.
- **Co wprowadza w błąd**: Fotorealizm, wskazówki dotyczące trendów mody.
- **Intencja**: Humorystyczna
- **Szkoda**: Podważa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

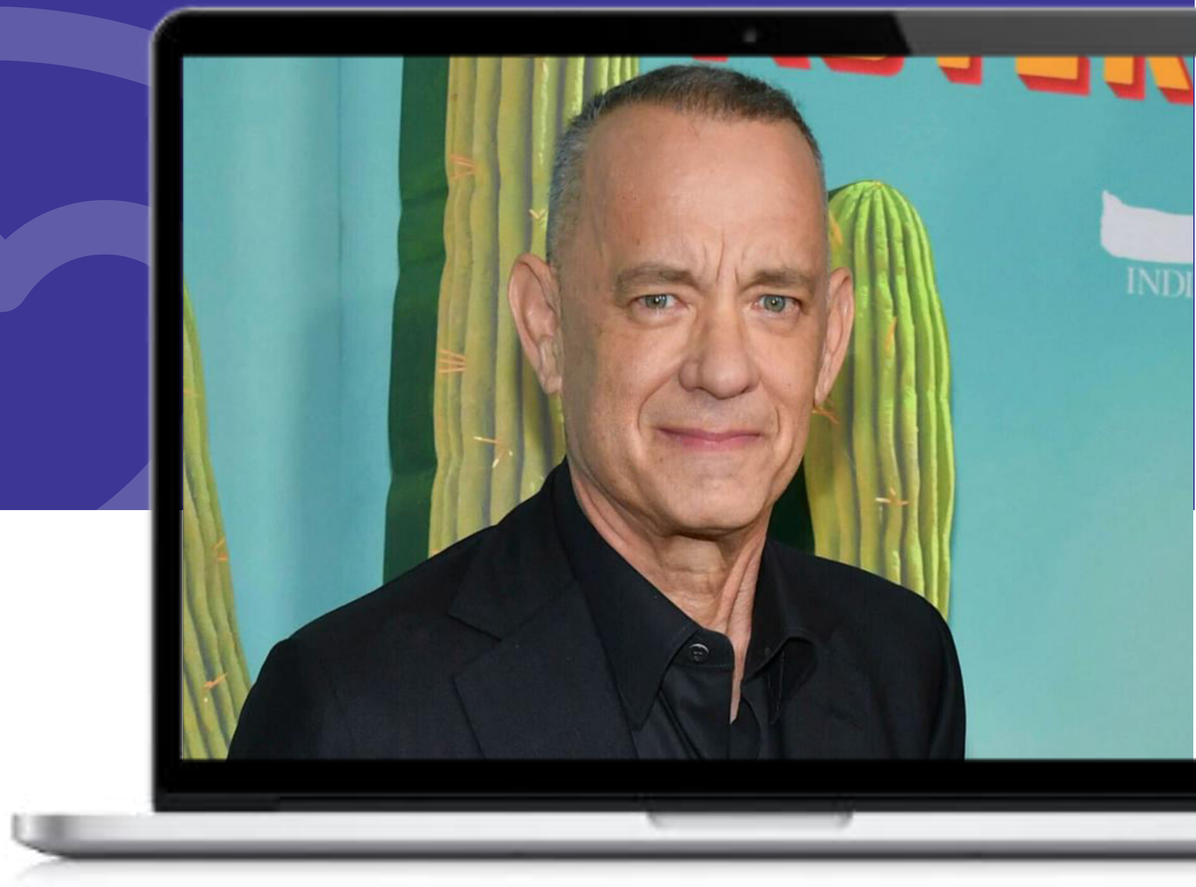
- **Co się wydarzyło**: Pracownik działu finansowego w Hongkongu padł ofiarą oszustwa, uczestnicząc w wieloosobowej, fałszywej rozmowie wideo z rzekomymi „współpracownikami”, w wyniku czego przekazał łącznie około 25 mln dolarów.
- **Co wprowadza w błąd**: Przekonujące klony nagrań wideo oraz głosów 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.

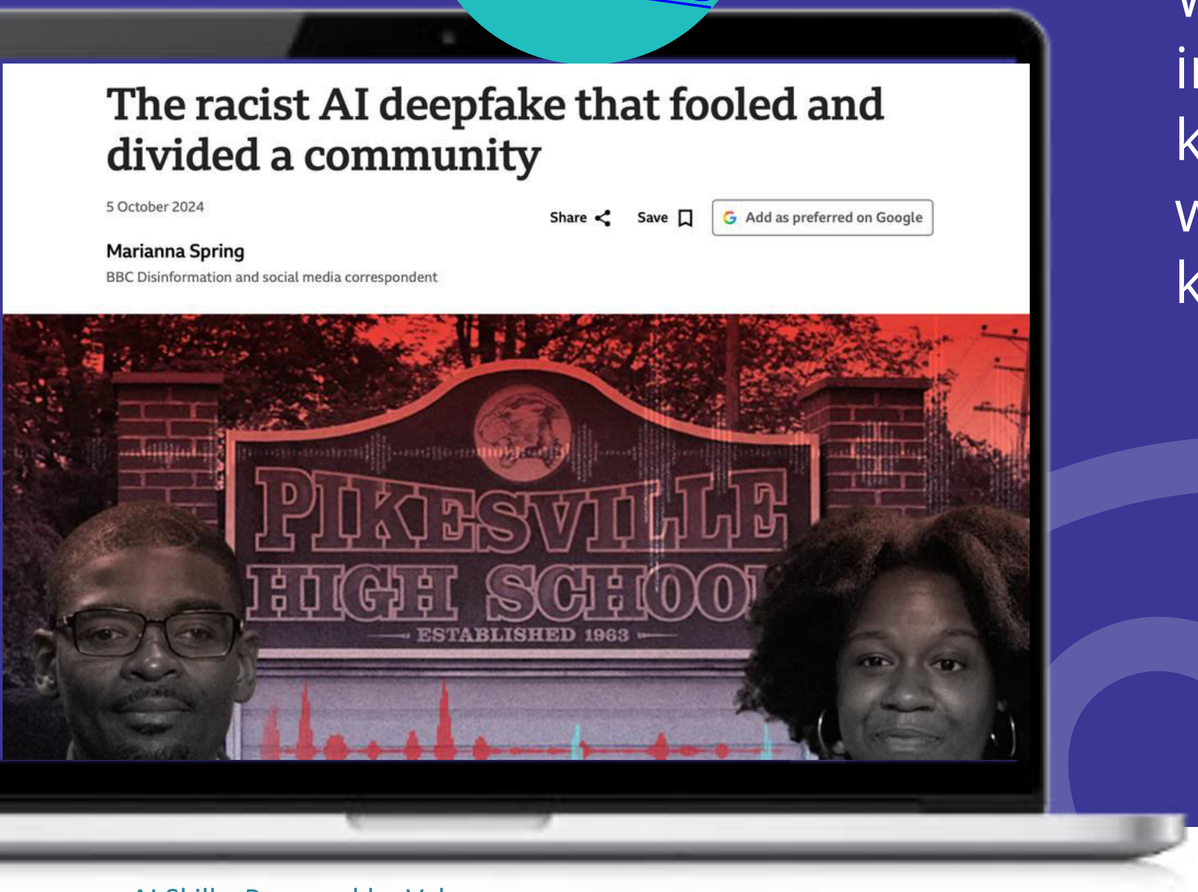
Co powinien uczynić wyborca po odebraniu takiego telefonu?

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

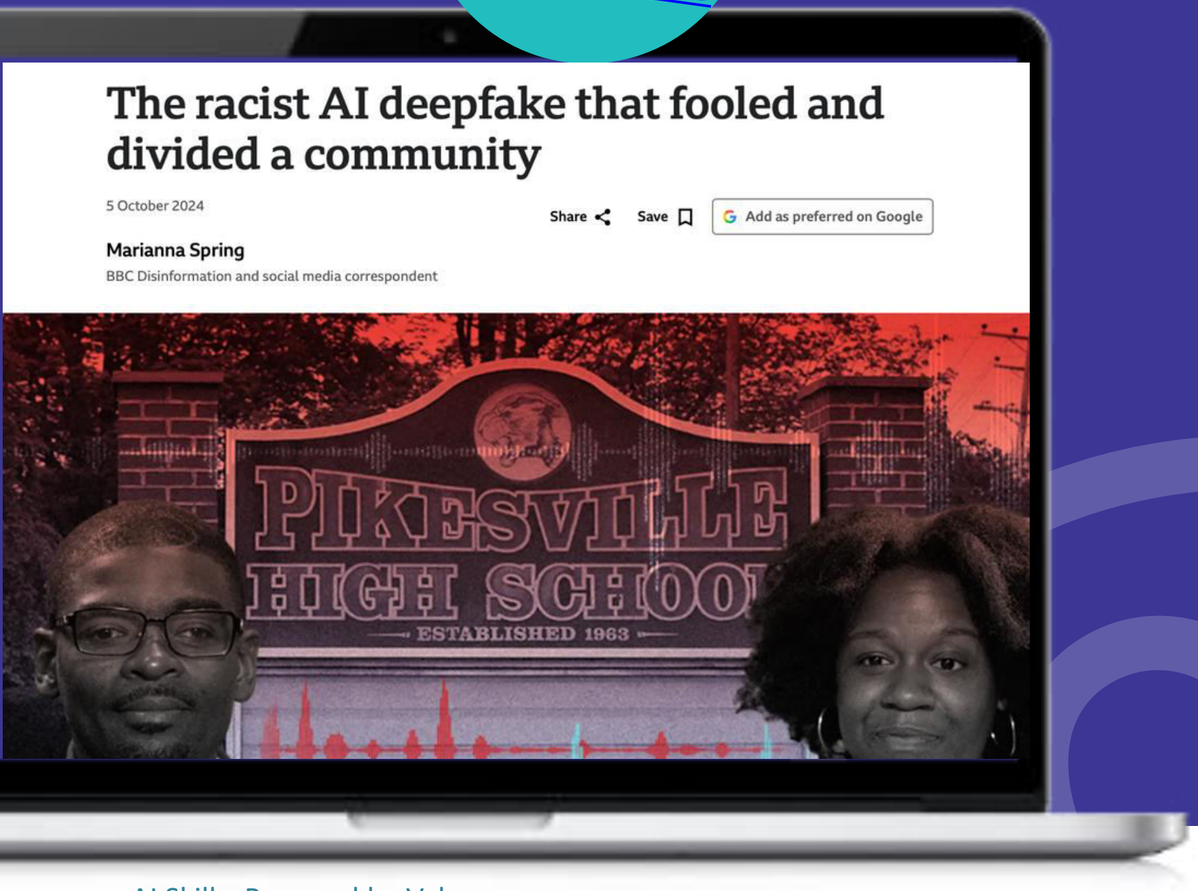


W miasteczku nieopodal Baltimore (USA) w internecie opublikowano nagranie audio, na którym słychać wypowiedzi dyrektora szkoły wygłaszającego rasistowskie i antysemickie komentarze.

- 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.

AI Skills, Powered by Values

values

www.valuesai.eu

MODUŁ 04

Niezawodna treść wspierana sztuczną inteligencją

Źródła, weryfikacja oraz kontrola halucynacji

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

01	Tworzenie rzetelnych treści w erze sztucznej inteligencji (10 min)	3
02	Zagrożenia związane z treściami sztucznej inteligencji (20 min)	8
03	Odpowiedzialne autorstwo (20 min)	15
04	Praca z materiałami źródłowymi (15 min)	23
05	Halucynacje sztucznej inteligencji (10 min)	36
06	Weryfikacja treści (10 minut)	45
07	Analiza danych z zastosowaniem sztucznej inteligencji (10 min)	51

Materiał dostępny na licencji Creative Commons CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

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

Tworzenie rzetelnych treści w erze sztucznej inteligencji

(10 minut)

Niezawodna zawartość AI

Tworzenie wiarygodnych treści w erze sztucznej inteligencji to nie tylko szybkie pisanie, ale przede wszystkim zapewnienie dokładności, przejrzystości i odpowiedzialności. Wykorzystanie sztucznej inteligencji nie obniża automatycznie jakości, lecz wymaga intensywniejszego nadzoru ze strony człowieka, szczególnie w kontekście faktów, danych, zdrowia, polityki czy kwestii społecznych.

AI Skills,
Powered
by Values
values

Niezawodna treść AI – fundamentalne zasady

- **Podawaj źródła.** Odwołuj się do badań, raportów, oficjalnych oświadczeń oraz oryginalnych danych.
- **Zweryfikuj przed publikacją.** Każdy element treści powinien być dokładnie sprawdzony pod względem faktów, danych liczbowych oraz kontekstu.
- **Ujawniaj zastosowanie sztucznej inteligencji.** Przejrzystość wzmacnia zaufanie odbiorców.
- **Pisz zwięźle i precyzyjnie.** Unikaj niejednoznacznych sformułowań, które mogą prowadzić do błędnej interpretacji.
- **Aktualizuj treść.** W przypadku wystąpienia nowych informacji lub wykrycia błędu, skoryguj materiał.
- **Zadbaj o wyraźne określenie autorstwa i odpowiedzialności.** Treść powinna mieć rozpoznawalnego autora lub redaktora.
- **Nie manipuluj odbiorcami.** Sztuczna inteligencja nie powinna być wykorzystywana do celowego wprowadzania ludzi w błąd, nawet jeśli przyczynia się to do zwiększenia zasięgu.

Niezawodna treść AI – co kształtuje zaufanie

Wiarygodne treści zazwyczaj wyjaśniają nie tylko co, ale również dlaczego i na jakiej podstawie, dlatego połączenie danych z komentarzem eksperta oraz praktycznymi przykładami sprawdza się doskonale.

- Daty wydania i aktualizacji,
- Linki do źródeł pierwotnych,
- Ekspert lub autor dysponujący odpowiednimi kwalifikacjami,
- Logiczna struktura tekstu.
- Wyraźne rozróżnienie pomiędzy faktem, opinią a interpretacją.
- Bez sensacji i języka nastawionego na przyciąganie kliknięć.

W praktyce wiarygodne treści w erze sztucznej inteligencji powstają, gdy sztuczna inteligencja wspiera pracę, ale to ludzie wciąż ponoszą odpowiedzialność za prawdziwość i sens przekazu.

h .

Dlaczego „wiarygodna treść” ma znaczenie w kontekście sztucznej inteligencji

Narzędzia AI są w stanie szybko generować treści, jednak każdy rezultat stanowi jedynie szkic: płynny tekst, wiarygodne dane i przekonujące podsumowania, które mogą okazać się błędne.

Dlaczego to ma znaczenie:

- **Po publikacji, niska jakość treści może zaszkodzić Twojej wiarygodności i wprowadzić odbiorców w błąd.**
- **Sztuczna inteligencja jest zdolna do generowania cytatów, statystyk i wypowiedzi, które wydają się autentyczne (halucynacje).**
- **„Niezawodny” oznacza coś więcej niż biegły – oznacza oparty na źródłach, zweryfikowany i odpowiedzialny.**

Zapytanie do ChatGPT o zdefiniowanie pojęcia skutkuje płynną odpowiedzią, jednak bez wskazania źródeł. To samo pytanie w NotebookLM, z załadowanymi akademickimi plikami PDF, generuje definicję z odniesieniami do rzeczywistych autorów.

02

Ryzyko związane z treściami sztucznej inteligencji

(20 minut)

Treści generowane przez sztuczną inteligencję niosą ze sobą trzy zagrożenia:

Halucynacje, brakujące źródła oraz wykrywanie sztucznej inteligencji

ludzie stosują sztuczną inteligencję nerowania treści, często skupiają się e na szybkości. W rzeczywistości generowane przez sztuczną encję wiążą się z trzema głównymi żeniami.

01

Ryzyko halucynacji (nieprawdziwe informacje):

Sztuczna inteligencja jest zdolna do generowania dat, statystyk, nazw i cytatów, które brzmią autentycznie, lecz nigdy nie miały miejsca.

02

Ryzyko źródłowe (niedostatek podstaw):

Ogólne modele, takie jak ChatGPT 4.0, generują spójny tekst bez odniesień do rzeczywistych publikacji. Bibliografia jest niejasna, niekompletna lub sfabrykowana.

03

Ryzyko identyfikacji (odciski palców AI):

Detektory odnoszą się do ogólnego, listowego, formułowego tekstu generowanego przez AI. Wynik bez edycji osiąga 100% punktów AI, natomiast wynik po edytowaniu spada poniżej 10%.

„Wygląda wiarygodnie”, lecz w rzeczywistości takie nie jest (mit kontra rzeczywistość)

Zwróć uwagę na:

- Idealnie sformatowane cytaty z nieprawidłowymi numerami stron
- Prawdziwe nazwiska autorów połączone z fikcyjnymi dziełami.
- Prawdopodobne statystyki bez potwierdzonego źródła
- Płynne definicje, które są niezgodne z aktualną literaturą.

Mit	***„Jeśli coś ma naukowy charakter, jest wiarygodne.”***
Fakt	Sztuczna inteligencja doskonale imituje styl akademicki, jednocześnie wciąż tworząc nieprawdziwe informacje.
Mit	***„Jeśli detektor zidentyfikuje, że tekst został napisany przez człowieka, jest bezpieczny.”***
Fakt	Detektory analizują jedynie styl. Edytowane dane wyjściowe AI mogą osiągnąć 92% wartości oryginalnej, a mimo to wciąż zawierać halucynacje.

Czym jest AI SLOP?

Błędy AI to niskiej jakości, wprowadzająca w błąd lub niedokładna treść tworzona przez sztuczną inteligencję.

Brzmi to pewnie, ale obejmuje:

- Sfabrykowane fakty
- Błędne numery lub daty
- Nieprawidłowe nazwy, cytaty lub stwierdzenia naukowe
- Powtarzalną, ogólną strukturaę
- „Gładki” tekst pozbawiony jakichkolwiek dowodów

Dlaczego to się dzieje?

01

Model przewiduje wzorce, a nie prawdę – generuje najbardziej prawdopodobną odpowiedź na podstawie danych, ale w rzeczywistości nie „wie”, czy jest ona prawdziwa.

02

Może uzupełniać luki fikcyjnymi szczegółami – w przypadku braku informacji model może dodać coś, co brzmi logicznie, ale niekoniecznie odzwierciedla rzeczywistość.

03

Informacje nie zawsze są aktualne – jeśli wiedza modelu jest ograniczona czasowo, może on przedstawiać nieaktualne fakty lub pomijać niedawne wydarzenia.

04

Może wystąpić błędne zrozumienie kontekstu – niejasne pytanie, skrót lub dwuznaczność mogą prowadzić do niewłaściwej odpowiedzi.

Dlaczego to się dzieje?

05

Brzmi pewnie, nawet gdy jest błędne – sformułowanie może wydawać się niezwykle wiarygodne, mimo że treść zawiera błędy.

06

Może łączyć rzeczywiste fakty z halucynacjami – niektóre elementy odpowiedzi mogą być poprawne, podczas gdy inne są wymyślone lub niepotwierdzone.

07

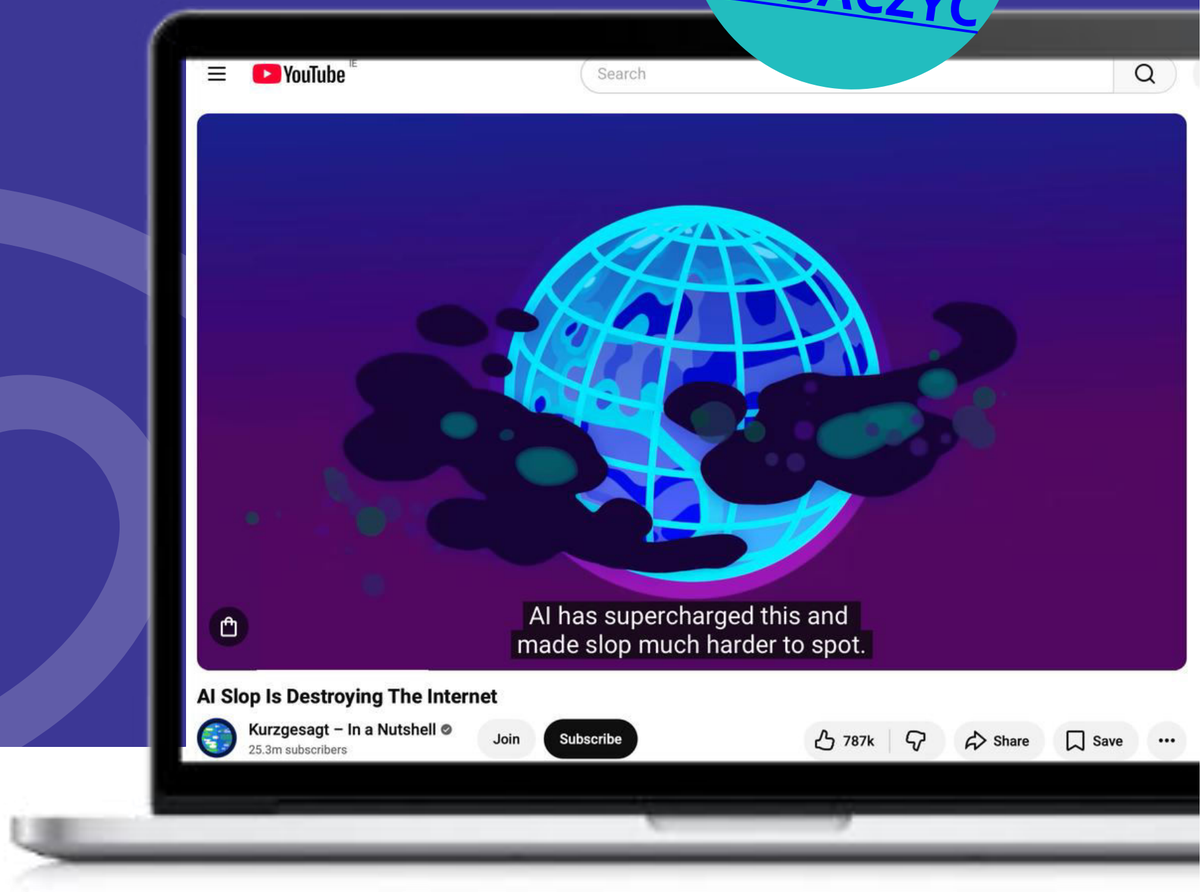
Zależy od jakości danych wejściowych – jeśli zapytanie jest niejasne, sprzeczne lub niekompletne, dane wyjściowe mogą być mniej precyzyjne.

latego weryfikacja staje się fundamentalną umiejętnością cyfrową.

Jak sztuczna inteligencja wpływa negatywnie na Internet

KLIKNIJ, ABY ZOBACZYĆ

Treści generowane przez sztuczną inteligencję zalewają internet – od filmów i muzyki po wiadomości i książki. Co więcej, generatywna sztuczna inteligencja sprawia, że fałszywe informacje wydają się bardziej przekonujące niż kiedykolwiek. Wkraczamy w nową erę przeciążenia informacyjnego, a odróżnianie prawdy od fikcji staje się coraz większym wyzwaniem.



03

Odpowiedzialne autorstwo

(10 minut)

Odpowiedzialny obiektyw.

Za każdym razem, gdy generujesz treści przy użyciu sztucznej inteligencji, zadaj sobie cztery pytania kontrolne:

Dokładność	Źródła	Oryginalność	Odpowiedzialność
Czy zweryfikowałem fakty, daty i cytaty, porównując je z rzetelnymi źródłami?	Czy mogę przedstawić autentyczne publikacje, na których opierają się wszystkie kluczowe twierdzenia?	Czy edytowałem projekt AI w sposób, który uniemożliwia jego wykrycie jako surowego wyniku AI?	Czy mogę przedstawić, w jaki sposób zastosowałem sztuczną inteligencję, co zmieniłem oraz dlaczego?

Rzetelna treść = inteligentne wykorzystanie sztucznej inteligencji z uwzględnieniem odpowiedzialności.

Odpowiedzialny obiektyw.

Bezpieczne korzystanie z AI	Prawa i zgody mają istotne znaczenie.	Sprawiedliwość stanowi integralny element bezpieczeństwa.	Istnieją bardziej bezpieczne przepływy pracy.
To, co wprowadzisz podczas przesyłania, ma kluczowe znaczenie.	Dane osobowe innych osób wymagają zgody oraz ochrony.	Dane stronnicze lub niskiej jakości mogą przynieść szkodę.	Często możesz uzyskać wsparcie sztucznej inteligencji, dzieląc się znacznie mniejszą ilością danych.

Zanim rozpoczniesz korzystanie ze sztucznej inteligencji: każdy rezultat stanowi wersję roboczą.

Wiele osób koncentruje się na szybkości sztucznej inteligencji. Wiarygodne treści zaczynają się wcześniej: od tego, co weryfikujesz, analizujesz i redagujesz.

W praktyce narzędzia sztucznej inteligencji generują:

- **Definicje oraz objaśnienia (często bez odniesień)**
- **Streszczenia prac (czasami fikcyjne lub zniekształcone)**
- **Listy odniesień (często fikcyjne)**
- **Analizy i interpretacje danych (bez metodologii)**

Zanim rozpoczniesz korzystanie ze sztucznej inteligencji:

każdy rezultat stanowi wersję roboczą.

Niezawodne zastosowanie sztucznej inteligencji rozpoczyna się od podstawowego nawyku: traktuj każdy wynik jako wersję roboczą do weryfikacji, a nie jako ostateczną odpowiedź.

AI Skills,
Powered
by Values
values

Wynik AI to zarys, a nie rzeczywistość.

Co uznaje się za treść wymagającą weryfikacji:

- Definicje oraz twierdzenia teoretyczne
- Statystyki, procenty oraz daty
- Nazwy autorów, książek oraz artykułów
- Analizy danych oraz wnioski

CO WOLNO

PAMIĘTAJ: Wszystkie fakty, cytaty i liczby pochodzące z sztucznej inteligencji powinny być zweryfikowane przed użyciem.

CZEGO NIE WOLNO

NIE zakładaj: „To brzmi profesjonalnie” lub „To musi być prawdą, skoro sztuczna inteligencja to stwierdziła”.

Lista
„Nigdy Nie Ufaj”

Jeżeli nie można zidentyfikować źródła, nie publiku

Nie publikuj bez weryfikacji:

- **Cytaty:** wszelkie odniesienia do książki, artykułu lub autora wskazanego przez sztuczną inteligencję.
- **Cytaty:** każdy „bezpośredni cytat” przypisany przez sztuczną inteligencję konkretnej osobie
- **Statystyka**: procenty, liczebność próby, wyniki badań, klasyfikacje
- **Fakty historyczne**: daty, wydarzenia, lokalizacje, nazwy aktów prawnych lub traktatów
- **Definicje i teorie**: stwierdzenia dotyczące tego, czym „jest” koncept lub kto go wprowadził.
- **Dane analizy numerycznej**: wykazy klientów, dane kadrowe, dokumenty wewnętrzne, osoby prywatne

Jak zwiększyć wiarygodność wyników sztucznej inteligencji (przed publikacją)

Przed publikacją wyników sztucznej inteligencji należy przeprowadzić następujące kontrole:

- Sprawdź cytaty: otwórz w Google Scholar, Scopus lub na stronie wydawcy.
- Sprawdź każdą statystykę przynajmniej w jednym niezależnym źródle.
- Uruchom ponownie prompt w zweryfikowanym narzędziu (np. NotebookLM z załadowanymi autentycznymi plikami PDF)
- Edytuj strukturę, zastąp ogólne frazy, wprowadź konkretne przykłady.

żeli nie jesteś w stanie zlokalizować źródła roszczenia → nie publikuj go.

04

Praca z materiałami źródłowymi

(15 minut)

Źródła dostarczone przez sztuczną inteligencję.

01

Korzystaj z rzetelnych źródeł – wybieraj materiały z uznawanych czasopism, instytucji publicznych, baz danych naukowych oraz organizacji eksperckich.

02

Rozróżniamy źródła naukowe, branżowe, medialne oraz opiniotwórcze – każdy z tych typów źródeł pełni odmienną funkcję i zapewnia różny poziom wiarygodności.

03

Weryfikacja dat publikacji – aktualność jest istotna, szczególnie w szybko zmieniających się obszarach, w których sztuczna inteligencja może generować nieaktualne dane.

04

Weryfikacja cytatów – upewnij się, że cytowany autor, tytuł, rok oraz przytoczone stwierdzenie rzeczywiście istnieją i są zgodne z treścią tekstu lub sztucznej inteligencji.

Źródła dostarczone przez sztuczną inteligencję.

05

Rozróżnienie między źródłami pierwotnymi a wtórnymi – źródło pierwotne dostarcza oryginalnego materiału, podczas gdy źródło wtórne analizuje lub interpretuje wcześniejsze prace.

06

Weryfikacja krzyżowa informacji – kluczowe twierdzenia powinny być potwierdzone w co najmniej dwóch lub trzech niezależnych źródłach.

07

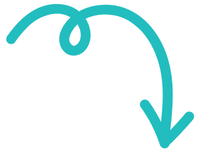
Ocena autora, wydawcy i celu – zawsze zastanów się, kto jest autorem tekstu, gdzie został opublikowany oraz czy jego celem jest informowanie, sprzedaż, perswazja czy komentarz.

Przykład podania przez sztuczną inteligencję błędnego źródła

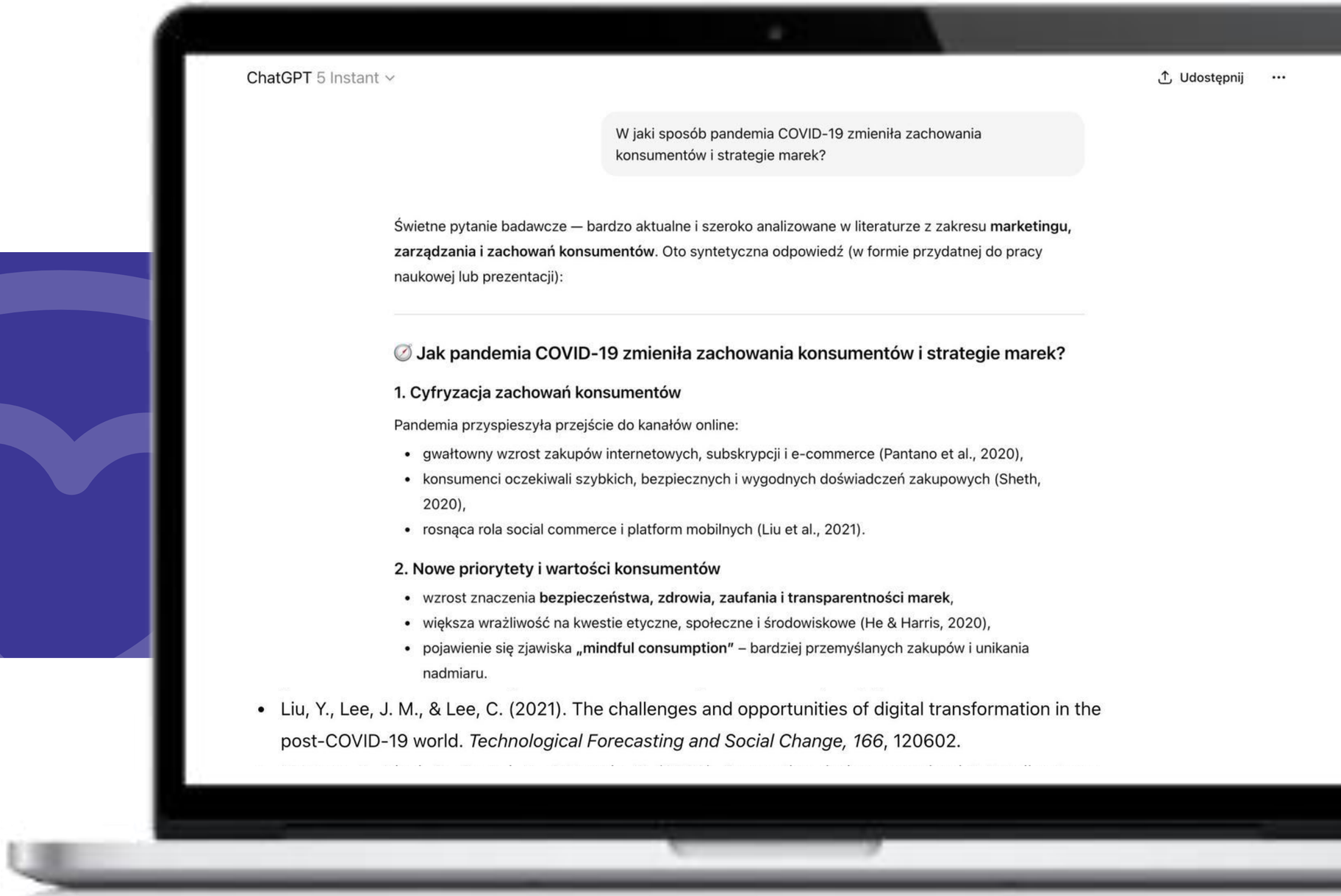
Ten slajd ilustruje rozmowę w ChatGPT 5 Instant, w której użytkownik pyta, w jaki sposób pandemia COVID-19 wpłynęła na zachowania konsumentów oraz strategie marek. Sztuczna inteligencja dostarcza zorganizowaną odpowiedź, odwołując się do źródeł akademickich (Pantano i in., 2020; Sheth, 2020; Liu i in., 2021; He i Harris, 2020), analizując digitalizację zachowań konsumenckich oraz nowe priorytety, takie jak bezpieczeństwo, zdrowie, zaufanie i „świadoma konsumpcja”.

Gdy użytkownik zweryfikował jedno cytowanie, odkrył, że artykuł „Wyzwania i możliwości transformacji cyfrowej w świecie po COVID-19” **nie został opublikowany** w czasopiśmie Technological Forecasting and Social Change, jak twierdzi AI, lecz w czasopiśmie IEEE Engineering Management Review, tom 48(3), s. 97–103.

Zobacz kolejny slajd



Przykład podania przez sztuczną inteligencję błędnego źródła

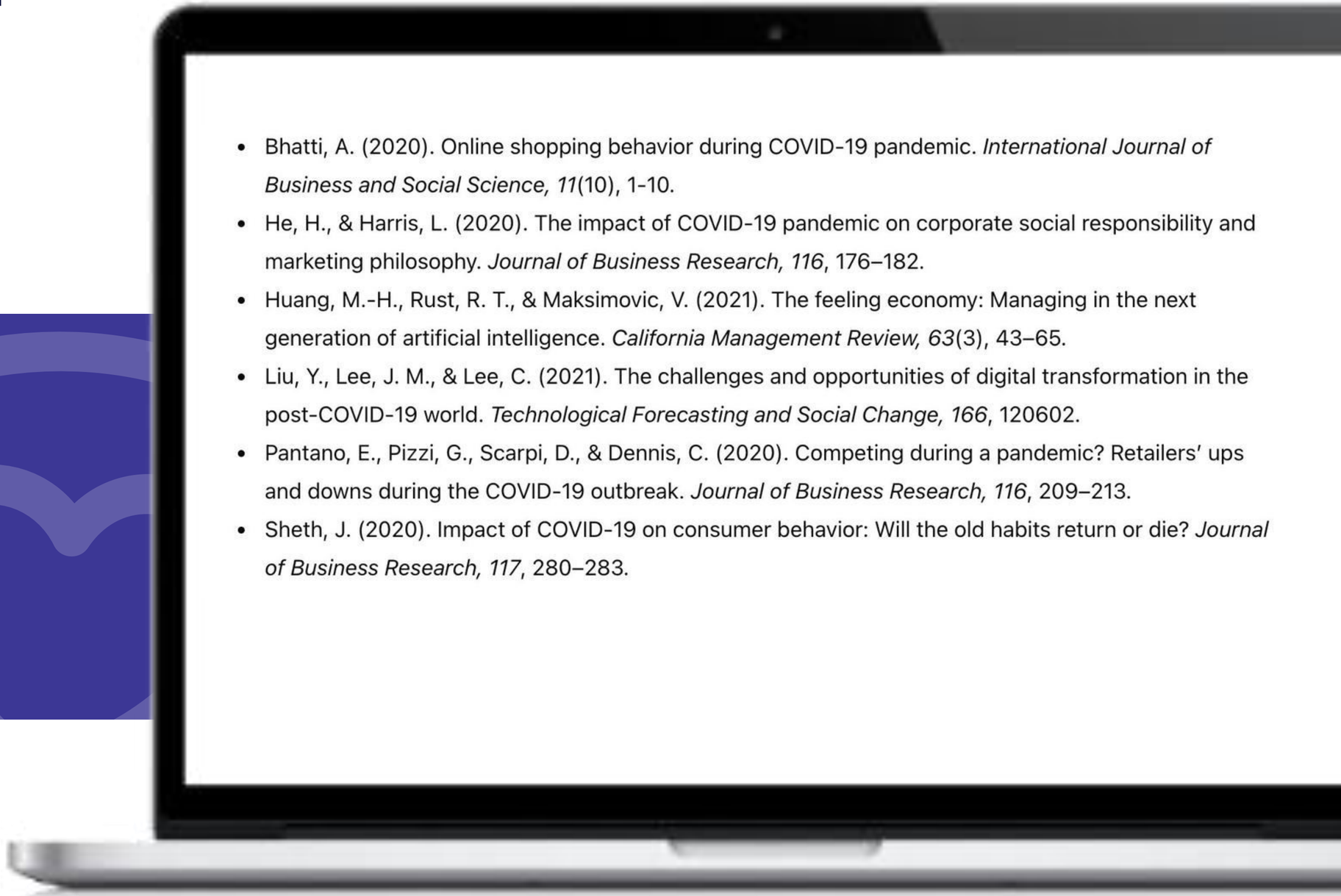


Przykład podania przez sztuczną inteligencję błędnego źródła. Wnioski

- Sztuczna inteligencja błędnie przypisała autentyczny artykuł niewłaściwemu czasopismu, mylnie cytując go jako opublikowany w Technological Forecasting and Social Change zamiast w IEEE Engineering Management Review.
- Wygenerowało to wiarygodny, aczkolwiek fałszywy odnośnik bibliograficzny, zawierający nazwę czasopisma oraz szczegóły publikacji, co doprowadziło do wprowadzenia w błąd w stylu cytowania APA.
- Sztuczna inteligencja nie była w stanie zweryfikować metadanych publikacji (czasopismo, wydawca, DOI), mimo że źródło było łatwe do potwierdzenia.
- Połączono autentyczny tytuł z niewłaściwym kontekstem publikacji, co stanowi typowy przykład halucynacji źródłowej.
- Błąd ten podważa naukową wiarygodność, co może skutkować nieprawidłowymi bibliografiami oraz problemami podczas recenzji eksperckich.

Zobacz kolejny slajd

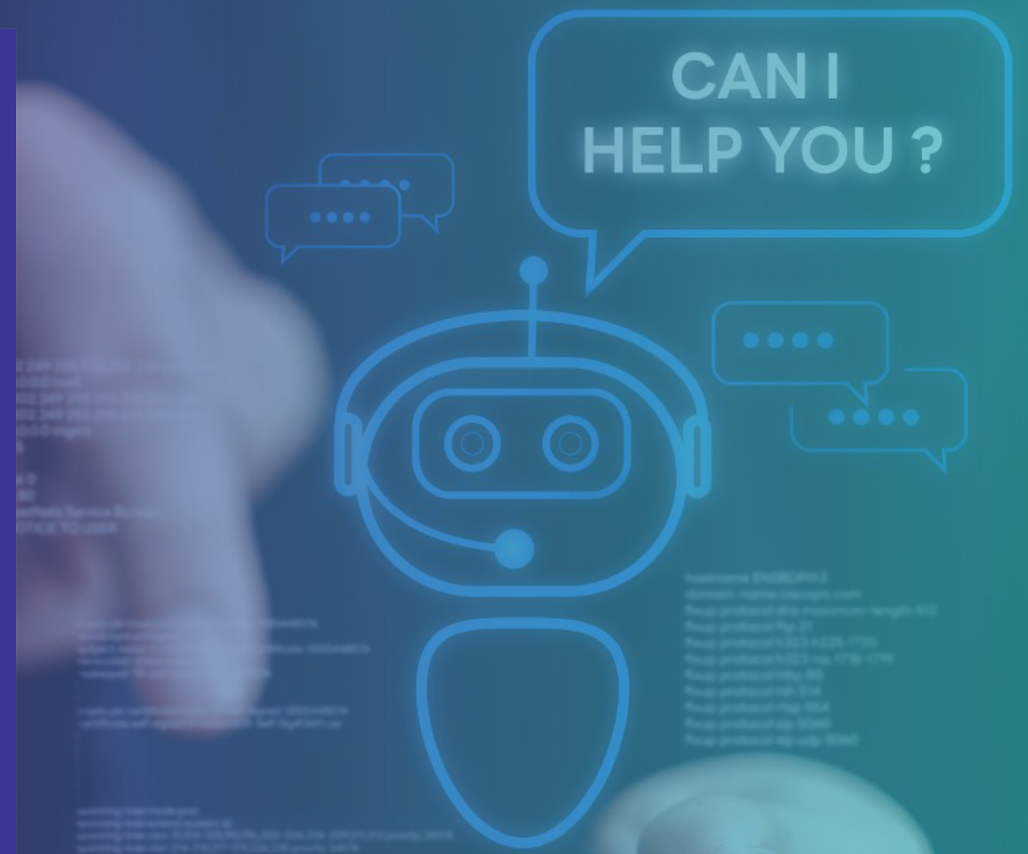
Przykład podania przez sztuczną inteligencję błędnego źródła. Wnioski



- Bhatti, A. (2020). Online shopping behavior during COVID-19 pandemic. *International Journal of Business and Social Science, 11*(10), 1-10.
- He, H., & Harris, L. (2020). The impact of COVID-19 pandemic on corporate social responsibility and marketing philosophy. *Journal of Business Research, 116*, 176–182.
- Huang, M.-H., Rust, R. T., & Maksimovic, V. (2021). The feeling economy: Managing in the next generation of artificial intelligence. *California Management Review, 63*(3), 43–65.
- Liu, Y., Lee, J. M., & Lee, C. (2021). The challenges and opportunities of digital transformation in the post-COVID-19 world. *Technological Forecasting and Social Change, 166*, 120602.
- Pantano, E., Pizzi, G., Scarpi, D., & Dennis, C. (2020). Competing during a pandemic? Retailers' ups and downs during the COVID-19 outbreak. *Journal of Business Research, 116*, 209–213.
- Sheth, J. (2020). Impact of COVID-19 on consumer behavior: Will the old habits return or die? *Journal of Business Research, 117*, 280–283.

Przykład podania przez sztuczną inteligencję nieprawidłowego źródła. Wnioski

Sztuczna inteligencja może skutecznie podsumowywać literaturę, ale wszystkie generowane przez nią źródła muszą zostać niezależnie zweryfikowane przed wykorzystaniem w pracy naukowej.



AI Skills, Powered by Values
values

Ćwiczenia

Poproś dowolny model AI o wygenerowanie jednostronicowego tekstu w formacie A4 na wybrany temat. Zachęć go, aby nie korzystał z internetu, lecz opierał się na publikacjach, takich jak artykuły, książki, raporty i inne podobne źródła.

- Utwórz wykaz źródeł.
- Następnie, zgodnie z najlepszymi praktykami, zweryfikuj wiarygodność podanych źródeł: czy rzeczywiście istnieją?
- Czy są one dostępne?
- Czy podają właściwych autorów?

Jeśli dostrzeżesz błąd sztucznej inteligencji, podziel się swoimi spostrzeżeniami z grupą oraz instruktorem.

Scenariusz: ChatGPT 4.0 vs NotebookLM

W analizie przypadku porównano dwa narzędzia przeznaczone do realizacji tego samego zadania:
„Generowanie definicji strategii innowacji z bibliografią”.

To samo polecenie wygenerowało dwa różne wyniki:

- ChatGPT 4.0: precyzyjna definicja, brak autentycznej bibliografii, zidentyfikowany jako 100% AI przez Originality.ai.
- NotebookLM (z autentycznymi plikami PDF): uzasadniona definicja opierająca się na rzeczywistych autorach (Porter, Janasz, Pomykalski, Freeman); 92% oceny oryginalnej.

Co to oznacza w praktyce?

- Bez solidnych podstaw, sztuczna inteligencja generuje płynne, lecz nieudokumentowane treści. Dysponując rzeczywistą bazą wiedzy, ten sam model wytwarza weryfikowalny tekst związany z konkretnymi publikacjami.

ChatGPT 4.0 a NotebookLM:

Źródła nie mają wyłącznie charakteru naukowego – to one kształtują odpowiedzialność za wyniki sztucznej inteligencji.

ChatGPT 4.0 (bez osadzenia):

- Płynny tekst, zwięzłe definicje, brak bibliografii
- Prawdopodobne przykłady przedsiębiorstw (Google, Tesla, Apple) nieposiadające żadnych „badań” w swoim dorobku.
- 100% wykryto jako wygenerowane przez sztuczną inteligencję przez Originality.ai

NotebookLM (z załadowanymi dokumentami PDF):

- Definicje oparte na teorii, odwołujące się do Portera, Janasza, Pomykalskiego i Freemana
- Prawdziwe tradycje badawcze: typologie innowacji, gospodarka sieciowa, kontekst małych i średnich przedsiębiorstw.
- Konkretna bibliografia w formacie APA, przygotowana do weryfikacji.
- Styl oceniony na 92% jako autentyczny przez Originality.ai

Źródła i prawa:

„Czy posiadam jakiekolwiek rzeczywiste źródło dla tego twierdzenia?”

Niezawodne zastosowanie sztucznej inteligencji to nie tylko fakty. To także kwestia autorstwa: kto jest twórcą oryginalnego pomysłu, czyje dzieło przytaczasz i w jaki sposób je przytaczasz.

Zazwyczaj można polegać na sztucznej inteligencji, jednak wyniki AI nie powinny być traktowane jako wiarygodne w następujących kwestiach:

- Dokładne statystyki, procenty, rozmiary próby
- Konkretne cytaty przypisywane wskazanym autorom
- Lata publikacji, numery stron, tytuły periodyków
- Definicje niszowych lub innowacyjnych koncepcji, które nie występują w literaturze głównego nurtu.
- Ogólne wyjaśnienia powszechnie znanych pojęć
- Edycja strukturalna (gramatyka, płynność, akapity)
- Podsumowanie tekstu, który sam dostarczyłeś sztucznej inteligencji.

Jak definiuje się „prawdziwe odniesienie” (i czym nie jest)?

Niezawodne zastosowanie sztucznej inteligencji to nie tylko fakty. To także kwestia autorstwa: kto jest twórcą oryginalnego pomysłu, czyje dzieło przytaczasz i w jaki sposób je przytaczasz.

Prawdziwe źródło to nie tylko nazwisko i rok. W przypadku cytatów generowanych przez sztuczną inteligencję źródło musi cechować:

- **Rzeczywistość**: publikacja rzeczywiście istnieje i jest dostępna w bazie danych.
- **Konkretność**: dosłowne dopasowanie żądania, a nie jedynie tematu
- **Dostępność**: możesz ją otworzyć i zapoznać się z odpowiednią sekcją.
- **Autorytatywność**: opublikowana przez rzetelne czasopismo, wydawcę lub instytucję
- **Ważne**: zgoda jest uzależniona od odbiorców oraz narzędzia.

Ktoś może zgodzić się przekazać ci pewne informacje, ale nie wyrazić zgody na ich udostępnienie narzędziu opartemu na sztucznej inteligencji lub na ich przechowywanie w systemie.

05

Halucynacje sztucznej inteligencji

(10 minut)

Czym są halucynacje sztucznej inteligencji?

Halucynacje sztucznej inteligencji to odpowiedzi generowane przez model, które brzmią wiarygodnie, lecz zawierają fałszywe, zmyślone lub wprowadzające w błąd informacje. Model „nie wie”, co jest prawdą – zazwyczaj przewiduje najbardziej prawdopodobny tekst, zamiast weryfikować rzeczywistość.

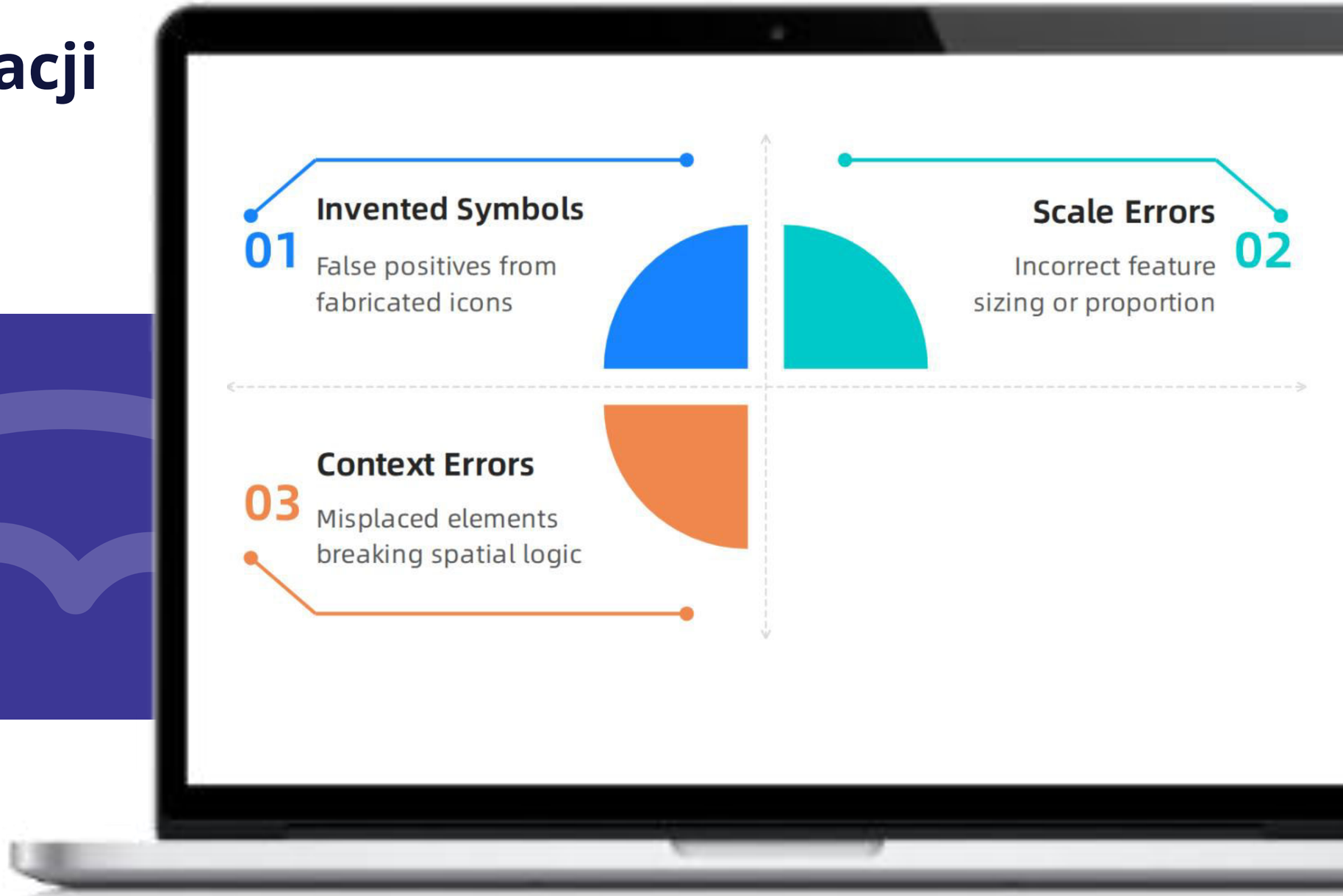
Halucynacje AI mają miejsce, gdy system tworzy treść, która wydaje się poprawna, lecz jest nieprawdziwa pod względem faktograficznym, na przykład zmyślony cytat, nieistniejące źródło, błędna data lub mylące wyjaśnienie. Nie są to halucynacje w sensie medycznym – to błędy w generowaniu informacji, a nie doświadczenia percepcyjne.

Mapa halucynacji sztucznej inteligencji

Jak identyfikować i naprawiać halucynacje sztucznej inteligencji w mapach symboli stopniowanych.

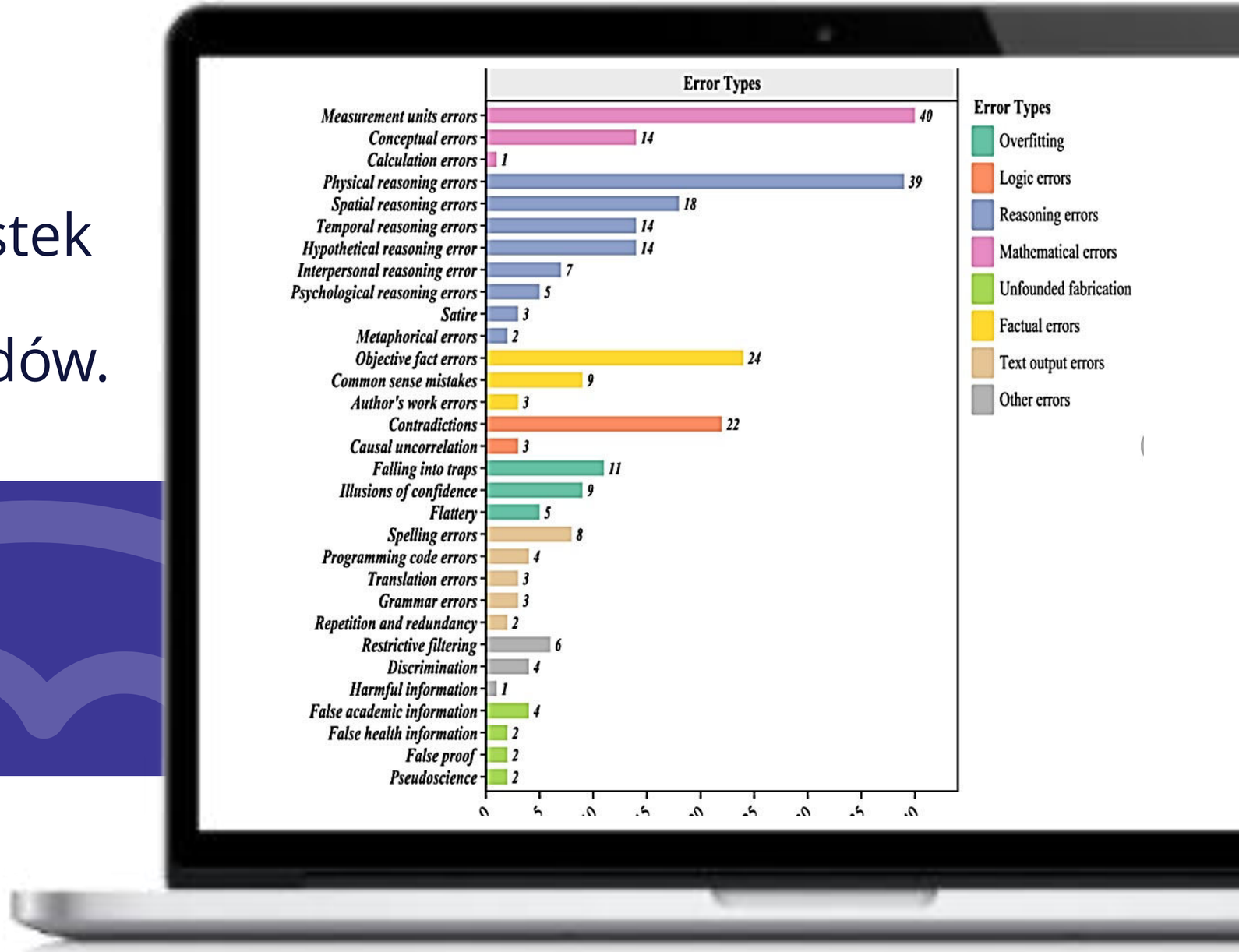
Aihallucinationreport.com.

https://aihallucinationreport.com/how-to-detect-and-fix-ai-hallucinations-in-graduated-symbol-maps/



Halucynacje sztucznej inteligencji

Statystyki dla 284 jednostek analizy. Różne kolory wskazują różne typy błędów.



Sun, Yujie, Sheng, Dongfang, Zhou, Zihan, Wu, Yifei. (2024). Halucynacja AI: w kierunku kompleksowej klasyfikacji zniekształconych informacji w treściach generowanych przez sztuczną inteligencję. Komunikacja w naukach humanistycznych i społecznych, 11. 10.1057/s41599-024-03811-x.

Rodzaje halucynacji sztucznej inteligencji

01

Halucynacje faktyczne – model przedstawia nieprawdziwe fakty, zdarzenia, osoby lub definicje.

02

Halucynacje źródłowe – sztuczna inteligencja tworzy fikcyjne źródła, publikacje, autorów, tytuły lub przypisuje sobie prawa do nieistniejących tekstów.

03

Halucynacje cytatowe – model przedstawia cytat, który w rzeczywistości nigdy nie został wypowiedziany lub błędnie przypisuje go danej osobie.

04

Halucynacje liczbowe – sztuczna inteligencja niepoprawnie interpretuje dane, myli procenty, daty, statystyki lub wartości tabelaryczne.

Rodzaje halucynacji sztucznej inteligencji

05

Halucynacje kontekstowe – model błędnie interpretuje pytanie lub wprowadza szczegóły, których użytkownik nie zamierzał uwzględnić.

06

Mieszane halucynacje – odpowiedź zawiera zarówno rzeczywiste fakty, jak i fikcyjne elementy, co utrudnia identyfikację błędu.

Przyczyny halucynacji sztucznej inteligencji

01

Model przewiduje wzorce, a nie prawdę – generuje najbardziej prawdopodobne słowa, zamiast weryfikować fakty.

02

Dane szkoleniowe mogą być niepełne lub błędne – jeśli zawierają luki, błędy lub sprzeczności, model jest w stanie je zrekonstruować.

03

Brak bieżących informacji – model może nie być świadomy najnowszych wydarzeń lub zmian.

04

Niejasny kontekst wskazówek – w przypadku niejednoznacznych lub skrótowych wskazówek model może błędnie interpretować intencje użytkownika.

Przyczyny halucynacji sztucznej inteligencji

05

Systemy nagradzają płynne odpowiedzi – modele są często trenowane w taki sposób, aby udzielały pewnych i płynnych odpowiedzi, nawet w sytuacjach, gdy występuje niepewność.

06

Brak wbudowanej funkcji weryfikacji faktów – bez dostępu do zewnętrznych źródeł model nie jest w stanie potwierdzić, czy jego odpowiedź jest prawdziwa.

Halucynacje pośrednie:
kiedy imiona i daty wydają się poprawne, ale w rzeczywistości takie nie są

Typowe halucynacje pośrednie:

- Prawdziwy autor + prawdziwe czasopismo + fikcyjny tytuł artykułu
- Prawdziwy dokument + nieprawidłowy rok, strona lub numer tomu
- Prawdziwa koncepcja + niewłaściwy autor („Porter stwierdził”, gdy był to Schumpeter)
- Rzeczywista statystyka + fikcyjny kontekst (prawidłowa liczba, błędne badanie)

Bezpieczniejsze opcje:

- **Wyszukaj**: otwórz Google Scholar i zweryfikuj, czy dany dokument istnieje.
- **Zastąp**: przytocz artykuł, który rzeczywiście przeczytałeś na ten sam temat
- **Zmiękcz**: zastosuj ogólne sformułowania, jeśli nie jesteś w stanie wskazać konkretnego źródła.
- **Usuń**: jeśli nie możesz zlokalizować źródła, usuń całe oświadczenie.

Sztuczna inteligencja może doświadczać halucynacji, nawet gdy imiona, daty i tytuły wydają się być poprawne. Schemat jest prawdopodobny, jednak związek z oryginalnym dziełem jest naruszony.

06

Weryfikacja treści

(10 minut)

Treści generowane przez sztuczną inteligencję mogą być użyteczne, jednak nie powinny być publikowane ani wykorzystywane bez uprzedniej weryfikacji.

Sprawdź twierdzenia, liczby, cytaty oraz źródła przed ich użyciem.

Potraktuj odpowiedź sztucznej inteligencji jako wstępny zarys, a nie jako ostateczną prawdę.

Udokumentuj, co zostało potwierdzone oraz na jakiej podstawie.

Krok 1:

Zidentyfikuj ryzyko oraz kluczowe roszczenia

Najpierw zidentyfikuj elementy odpowiedzi AI, które wymagają obowiązkowego sprawdzenia: fakty, daty, nazwy, liczby, cytaty, definicje oraz cytowane źródła. Im większy wpływ na zdrowie, prawo, edukację, finanse lub reputację, tym bardziej rygorystyczna powinna być weryfikacja.

Które elementy mogą spowodować szkody, jeśli są nieprawidłowe?

Czy sztuczna inteligencja dostarczyła autentyczne źródła, czy jedynie brzmi przekonująco?

Co w tej odpowiedzi stanowi fakt, a co jest interpretacją?

Krok 2:

Weryfikacja faktów i źródeł

Każde istotne stwierdzenie powinno być weryfikowane na podstawie wiarygodnych źródeł zewnętrznych, najlepiej pierwotnych: artykułów naukowych, raportów, stron internetowych instytucji publicznych oraz renomowanych publikacji branżowych. Nigdy nie zakładaj, że źródło cytowane przez sztuczną inteligencję rzeczywiście istnieje — zweryfikuj autora, tytuł, rok, dostępność oraz to, czy źródło rzeczywiście zawiera podane informacje.

Potwierdź informację w co najmniej dwóch niezależnych źródłach.

Otwórz oryginalne źródło, a nie jedynie streszczenie lub powtórne opublikowanie.

Sprawdź datę publikacji oraz aktualność informacji.

Krok 3:

Oceń jakość, język oraz kontekst

Po weryfikacji faktów oceń, czy tekst jest logiczny, klarowny, zrównoważony i adekwatny do zamierzonego celu. Sztuczna inteligencja potrafi generować treści, które są formalnie poprawne, lecz słabo osadzone w kontekście, nadmiernie pewne siebie, stronnicze lub stylistycznie nieodpowiednie.

W tekście fakty są pomieszane z opiniami.

Zawiera stronniczość, nadmierne uproszczenia oraz zbyt daleko idące uogólnienia.

Ton i styl są dostosowane do odbiorców oraz celu publikacji.

Edytuj, podejmuj decyzje i dokumentuj

Na koniec popraw błędy, usuń niejasne fragmenty i zdecyduj, czy treść jest odpowiednia do użytku. Dobrą praktyką jest odnotowanie, co zostało zmienione, co odrzucone oraz które źródła potwierdziły ostateczną wersję tekstu. Decyzja końcowa:

1

Publikuj

Jeśli treść została zweryfikowana i skorygowana.

2

Wstrzymaj się

Jeśli niektóre informacje pozostają niejasne.

3

Odrzuć

Jeśli treść zawiera zbyt wiele błędów lub brakuje weryfikacji źródła.

07

Analiza danych z zastosowaniem sztucznej inteligencji

(10 minut)

Gdy analizujesz dane przy użyciu sztucznej inteligencji, walidacja staje się kluczowym zadaniem.

Gdy zwrócisz się do sztucznej inteligencji o analizę danych, system stworzy spójną narrację. W zależności od polecenia i zastosowanego narzędzia, rezultat może również:

- Pominąć metodologię (brak odniesienia do wielkości próby, metody lub założeń)
- Uśrednić wartości
- Uogólnić na podstawie ograniczonej próby (np. n=10 respondentów)
- Wyciągać konkretne wnioski (gdy dane potwierdzają jedynie wnioski wstępne)
- Błąd potwierdzenia: sztuczna inteligencja może powtarzać treść komunikatu.

Gdy analizujesz dane przy użyciu sztucznej inteligencji, walidacja staje się kluczowym zadaniem.

Dlaczego walidacja analizy jest istotna:

- Wykrywasz błędy zanim trafią do raportu lub publikacji.
- Potrafisz rozróżnić solidne ustalenia od przekazu informacyjnego, który wydaje się wiarygodny.
- Unikasz podejmowania decyzji na podstawie próbek, które nie są reprezentatywne.
- Zachowujesz autorstwo analizy, a nie modelu.

Drabina analizy SAFE: wybierz najbardziej weryfikowalny poziom

Analizując dane przy użyciu sztucznej inteligencji, nie rozpoczynaj od najwyższego poziomu. Wybierz najniższy poziom, który nadal odpowiada na Twoje pytanie.

Zasada zatrzymania: Jeśli nie jesteś w stanie usunąć wrażliwych lub poufnych elementów → nie przesyłaj.

POZIOM 1

(Najbardziej wiarygodny): Zadaj pytanie ogólne (bez danych osobowych).

„Wyjaśnij, w jakich okolicznościach test t jest stosowny, a kiedy należy zastosować test U Manna-Whitneya."

POZIOM 2

Użyj anonimowego podsumowania (role, bez identyfikatorów).

„Średnia = 4,2, SD = 1,1, n = 120 – co to implikuje w kontekście trendu?"

POZIOM 3

Używaj ograniczonych, zanonimizowanych danych wyłącznie w przypadku konieczności.

Przykład: „Oto zestaw danych. Oblicz X, a następnie ja porównam to z moimi obliczeniami."

Wnioski z ograniczonych danych

Niepewny rezultat sztucznej inteligencji (realistyczny):

„Sztuczna inteligencja przeprowadziła analizę sentymentu na podstawie 200 tweetów dotyczących naszej marki. 67% z nich miało charakter pozytywny. W związku z tym sentyment wobec marki jest zdecydowanie pozytywny.”

Co jest niepewne?

- „67%” bez jakiejkolwiek metodologii, modelu i walidacji
- 200 tweetów to niewielki, prawdopodobnie subiektywny fragment.
- „zdecydowanie pozytywny” to przesada w odniesieniu do tego, co wskazuje liczba.

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.

AI Skills, Powered by Values

values

www.valuesai.eu

MODUŁ 05

Prywatność danych oraz bezpieczne użytkowanie narzędzi AI

Udostępniaj mniej. Chroń więcej.
Korzystaj z sztucznej inteligencji
w sposób odpowiedzialny.

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

01	Dlaczego bezpieczeństwo danych jest istotne w kontekście sztucznej inteligencji (15 min)	3
02	Jak bezpiecznie dzielić się danymi? (15 min)	10
03	Prywatność, prawa oraz zgoda (15 min)	16
04	Wybór bardziej bezpiecznych przepływów pracy związanych ze sztuczną inteligencją (25 min)	23
05	Niektóre dane (15 min)	33

Materiał dostępny na licencji Creative Commons CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

y Values

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

Dlaczego bezpieczeństwo danych jest istotne w kontekście sztucznej inteligencji

(15 minut)

Dane oraz sztuczna inteligencja

Udostępnianie danych sztucznej inteligencji w miejscu pracy wiąże się z ryzykiem, ponieważ może prowadzić do ujawnienia informacji, naruszeń przepisów oraz szkód wizerunkowych lub operacyjnych, jeśli dane wejściowe zostaną zapisane, ponownie wykorzystane lub wyciekają do modeli innych firm.

AI Skills, Powered by Values

values

Kluczowe zagrożenia

- **Ujawnienie danych wrażliwych**: monity lub pliki przesyłane do platform AI mogą być przechowywane lub pojawić się w wynikach modelu, co może prowadzić do ujawnienia własności intelektualnej, danych osobowych klientów lub zapisów finansowych.
- **Shadow AI i niekontrolowane wykorzystanie**: pracownicy korzystający z nieautoryzowanych narzędzi AI tworzą martwe strefy, w których dane wymykają się spod kontroli korporacyjnej, a dostawcy mogą przechowywać lub ponownie wykorzystywać dane wejściowe.
- **Naruszenia przepisów prawnych i zgodności**: korzystanie z zewnętrznych usług AI bez odpowiednich umów lub porozumień dotyczących przetwarzania danych może prowadzić do naruszenia przepisów o ochronie prywatności i skutkować sankcjami.
- **Brak przejrzystości i możliwości audytu**: wielu dostawców sztucznej inteligencji nie ujawnia, w jaki sposób przechowywane są dane wejściowe, gdzie są przetwarzane ani czy są wykorzystywane do trenowania modeli, co utrudnia ocenę ryzyka.
- **Rozszerzona powierzchnia ataku**: integracja sztucznej inteligencji zwiększa podatność techniczną (interfejsy API, wtyczki, punkty końcowe), którą mogą wykorzystać atakujący do uzyskania dostępu do danych.

Trzy negatywne konsekwencje: prywatność, niesprawiedliwość i iluzoryczne zaufanie

01

Ryzyko naruszenia prywatności (eksponowanie):

Nadmierne udostępnianie informacji może prowadzić do ujawnienia danych osobowych, poufnych lub pochodzących od osób trzecich, czasami nawet w sposób pośredni.

02

Ryzyko integralności (szkoda)

Jeżeli dane są stronnicze, niekompletne lub zawierają niesprawiedliwe etykiety, wyniki sztucznej inteligencji mogą potęgować stereotypy lub dyskryminować określone grupy.

03

Ryzyko niezawodności (błędne zaufanie):

Sztuczna inteligencja potrafi generować płynne odpowiedzi, które brzmią poprawnie, nawet gdy dane wejściowe są słabe, pozbawione kontekstu lub mylące.

CO WOLNO

- **Zapytaj**: Czego brakuje? Kogo brakuje? Czy ten wynik może zaszkodzić komuś?
- **Kontrola użytkowania**: Consent Gate + Fair Data Check + SAFE Workflow Ladder

CZEGO NIE WOLNO

- **Traktuj wyniki sztucznej inteligencji jako obiektywną prawdę.**
- **Wykorzystuj wyniki sztucznej inteligencji do oceny ludzi lub podejmowania decyzji bez uwzględnienia kontekstu.**

Soczewka wartości.

Bezpieczne dane to mądre wybory oparte na wartościach.

Za każdym razem, gdy korzystasz ze sztucznej inteligencji opartej na danych, zadaj sobie cztery pytania dotyczące wartości:

1. **Decyzyjność**: Czy podejmuję świadomy wybór, czy klikam automatycznie?
2. **Prywatność**: Czy udostępniam więcej informacji (o sobie lub innych), niż jest to konieczne?
3. **Uczciwość**: Czy wykorzystanie moich danych może zaszkodzić komuś lub wyrządzić szkody?
4. **Przejrzystość**: Czy mogę przedstawić, jakie dane zostały wykorzystane i z jakiego powodu?

Szczególne kategorie danych osobowych.

Przy udostępnianiu tych danych systemom sztucznej inteligencji należy zachować szczególną ostrożność.

1. Dane osobowe
2. Dane poufne
3. Informacje zdrowotne
4. Dane finansowe
5. Dane do logowania
6. Dane Klienta
7. Wewnętrzne dokumenty administracyjne
8. Nieopublikowane wyniki badań
9. Dane objęte tajemnicą handlową

02

Jak bezpiecznie dzielić się danymi?

(15 minut)

Zanim rozpoczniesz korzystanie ze sztucznej inteligencji: Twoje dane wejściowe stają się danymi.

Wiele osób koncentruje się na tym, na jakie pytania odpowiada sztuczna inteligencja.

Bezpieczne dane zaczynają się na wcześniejszym etapie:
zależą od tego co wpisujesz, wklejasz lub przesyłasz.

W praktyce narzędzia sztucznej inteligencji mogą współpracować z:

- tekstem (monity, wiadomości, skopiowana treść)
- plikami (dokumenty, arkusze kalkulacyjne, pliki PDF)
- obrazami (fotografie, zrzuty ekranu)
- kontekstem (szczegóły umożliwiające wyciąganie wniosków)

Zanim rozpoczniesz korzystanie ze sztucznej inteligencji: Twoje dane wejściowe stają się danymi.

Jeśli udostępnisz zbyt wiele, możesz narazić się na ryzyko, które jest trudne do odwrócenia: zagrożenie dla prywatności, wyrządzenie szkody innym oraz publikowanie nieuczciwych lub wprowadzających w błąd treści.

Bezpieczne korzystanie ze sztucznej inteligencji rozpoczyna się od podstawowego nawyku: udostępniaj jedynie niezbędne informacje lub w ogóle ich nie udostępniaj.

Co wolno

- **Pamiętaj**: Wszystko, co wpisujesz, wklejasz lub przesyłasz, stanowi dane, którymi się dzielisz.

Czego nie wolno

- **NIE ZAKŁADAJ:** „To jedynie tekst” lub „To tylko szybka wiadomość”.

„Wygląda na nieszkodliwe”, ale takie nie jest (mit a rzeczywistość)

Mit

„Będzie bezpieczniej, jeśli usunę nazwę.”

Fakt

Ludzi można zidentyfikować pośrednio za pomocą kombinacji takich jak:

- wiek + miasto + szkoła/staż
- unikatowe szczegóły wydarzenia
- zrzuty ekranu z wyraźnymi, drobnymi wskazówkami
- wzorce lokalizacji i czasu

Mit

„To tylko zrzut ekrany, jest bezpieczny.”

Fakt

Zrzuty ekranu często zawierają imiona, twarze, dane kontaktowe lub kontekst, który może umknąć Twojej uwadze.

Zadanie: Zidentyfikuj sygnały ostrzegawcze

Zadanie (60–90 sekund):

Oznacz każdą pozycję jako Nigdy / Unikaj / Zwykle w Porządku:

- Kod do resetowania hasła
- CV z kompletnym adresem i numerem telefonu
- Zrzut ekranu czatu z widocznymi nazwami użytkowników
- Pytanie ogólne: „Jak stworzyć CV?”
- Lista szkół zawierająca nazwy oraz adresy e-mail.
- Podsumowanie ankiety zawierające jedynie sumy (bez nazwisk)

Nigdy nie wiesz.

- Nie zawsze jest oczywiste, gdzie odbywa się przetwarzanie danych.
- Dane mogą być archiwizowane w historii rozmów.
- Mogą być wykorzystywane do doskonalenia usług, w zależności od ustawień oraz warunków korzystania z usługi.
- Użytkownik może nieumyślnie ujawnić poufne informacje.

03

Prywatność, prawa oraz zgoda

(15 minut)

Zdecyduj, zanim podzielisz się danymi o innych osobach.

Używaj tej metody zawsze, gdy Twoje dane wejściowe zawierają informacje pochodzące od innych osób (wiadomości, zdjęcia, listy, opowieści, dokumenty).

Zadaj trzy pytania:

- **Prawo/pozwolenie**: Czy posiadam pozwolenie (lub wyraźne prawo) na wykorzystanie tych danych przez sztuczną inteligencję?
- **Ryzyko**: Czy może to kogoś zidentyfikować, zawstydzić lub skrzywdzić w chwili obecnej lub w przyszłości?
- **Bezpieczniejsza alternatywa**: Czy mogę zrealizować cel, korzystając z mniejszej ilości danych (lub nie wykorzystując żadnych danych osobowych)?

Reguła

Jeśli na jakiekolwiek pytanie odpowiesz „Nie” lub „Nie mam pewności”, nie przesyłaj danych. Wybierz bezpieczniejszy sposób pracy (np. zadaj ogólne pytanie bez ujawniania danych osobowych). „Nie mam pewności” to sygnał ostrzegawczy.

Zgoda oraz dane innych osób: na czym polega odpowiedzialne wykorzystanie danych.

Prywatność i prawa to nie tylko zagadnienia techniczne – dotyczą one szacunku oraz zapobiegania szkodom.

CO WOLNO

Dozwolone (bezpieczniejsze opcje):

- Poproś o zgodę i uszanuj odpowiedź „nie”.
- Zastąp nazwy rolami: Osoba A / przyjaciel / nauczyciel / kolega.
- Usuń szczegóły unikalne (konkretne lokalizacje, daty, identyfikatory).
- Zadaj ogólne pytanie: „W jaki sposób mogę wesprzeć przyjaciela w trudnej sytuacji?” (bez osobistej historii)

CZEGO NIE WOLNO

Niedozwolone (znaczne ryzyko):

- Wklejanie prywatnych wiadomości innej osoby do AI „w celu uzyskania porady”.
- Prześlij listy grupowe zawierające nazwy i adresy e-mail w celu ich „uporządkowania”.
- Dziel się osobistymi historiami o innych, nie zadając pytań.

Szybkie ćwiczenie: czy reguła zgody jest spełniona?

Instrukcje

W każdym scenariuszu podejmij decyzję: spełniona (wystarczająco bezpiecznie), niespełniona (wybrać bezpieczniejszą opcję). Następnie podaj jeden krótki powód.

- „Podsumuj list z opinią mojego nauczyciela, zawierający imię i nazwisko nauczyciela oraz nazwę szkoły.”
- „Proszę o pomoc w napisaniu przeprosin. Opiszę sytuację, nie wklejając całego czatu.”
- „Przeanalizuj listę wolontariuszy, która zawiera imiona, nazwiska oraz numery telefonów, aby zidentyfikować osoby „niegodne zaufania”.

Czy mam prawo do korzystania z tych danych?

Bezpieczne korzystanie ze sztucznej inteligencji to nie tylko kwestia prywatności, ale także praw: czy masz prawo do używania, udostępniania lub przetwarzania danych, zwłaszcza gdy należą one do innej osoby lub organizacji.

Zazwyczaj nie masz prawa do dzielenia się z sztuczną inteligencją:

- Prywatnymi wiadomościami innych użytkowników (nawet jeśli uczestniczysz w czacie)
- Listami osób (listy klasowe, listy wolontariuszy, listy klientów)
- Dokumentami, które nie są Twoją własnością (umowy, notatki działu HR, wewnętrzne e-maile)
- Osobistymi historiami, które zostały Ci powierzone w zaufaniu (zdrowie, relacje, problemy rodzinne)

Zazwyczaj przysługuje Ci prawo do korzystania z:

- Twojego osobisty tekstu, który nie zawiera żadnych poufnych informacji.
- Pytań ogólnych bez informacji osobowych
- Anonimizowanych przykładów, które samodzielnie stworzyłeś (fikcyjne lub zanonimizowane)

Zgoda: jak definiuje się „prawdziwą zgodę” (a czym nie jest?)

Zgoda to nie tylko „ktoś kiedyś powiedział tak”. W kontekście sztucznej inteligencji, zgoda:

- **Jest wydawana przez osobę doinformowaną**: osoba jest świadoma, co zostanie udostępnione oraz z jakiego powodu.
- **Jest konkretna**: zgoda na określony cel, a nie „cokolwiek zechcesz”
- **Jest dobrowolna**: bez przymusu, bez obaw o konsekwencje
- **Jest odwoływalna**: osoba ma możliwość zmiany zdania.
- **Ważne**: zgoda jest uzależniona od odbiorców oraz narzędzia. Ktoś może zgodzić się przekazać ci pewne informacje, ale nie wyrazić zgody na ich udostępnienie narzędziu opartemu na sztucznej inteligencji lub na ich przechowywanie w systemie.

Identyfikacja pośrednia: Kiedy „usunięcie nazw” okazuje się niewystarczające?

Ludzi można zidentyfikować nawet bez imion, jeśli szczegóły umożliwiają innym odgadnięcie, kim są.

Typowe identyfikatory pośrednie:

- Dokładny wiek, precyzyjna lokalizacja, szkoła/pracodawca, konkretne daty oraz wydarzenia związane z daną rolą („jedyna osoba, która…”)
- Kombinacje (wiek + miasto + pasja + okoliczność)
- Zrzuty ekranu z wyraźnymi wskazówkami (zdjęcia profilowe, nazwy grup)

Bezpieczniejsze opcje (wybierz jedną):

- **Streszczenie**: przedstaw sytuację bez ujawniania szczególnych detali.
- **Uogólnienie**: zakresy zastosowania (grupa wiekowa), rozległe obszary (region)
- **Oparcie na rolach**: Osoba A / Osoba B zamiast rzeczywistych tożsamości
- **Nie wysyłaj**: jeśli nie jesteś w stanie usunąć kontekstu identyfikującego

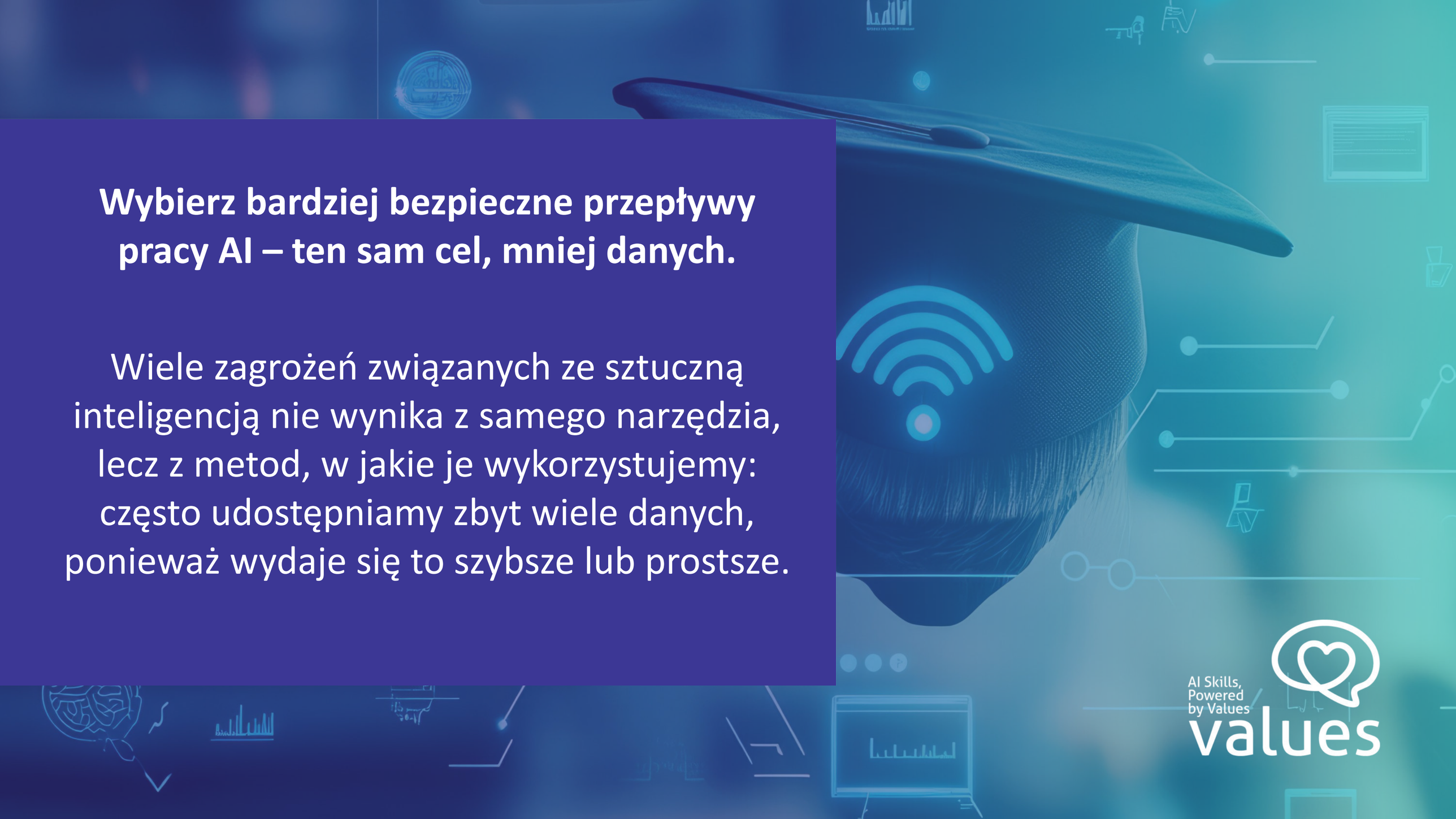
04

Wybór bardziej bezpiecznych przepływów pracy AI

(25 minut)

Wybierz bardziej bezpieczne przepływy pracy AI – ten sam cel, mniej danych.

Wiele zagrożeń związanych ze sztuczną inteligencją nie wynika z samego narzędzia, lecz z metod, w jakie je wykorzystujemy: często udostępniamy zbyt wiele danych, ponieważ wydaje się to szybsze lub prostsze.



AI Skills, Powered by Values

values

Chcesz wybrać bardziej bezpieczne przepływy pracy AI – ten sam cel, mniej danych.

Gdy udostępniasz dane sztucznej inteligencji, mogą być one przetwarzane, przechowywane i powielane.

Kiedy wprowadzasz, wklejasz lub przesyłasz dane do narzędzia AI, system musi je przetworzyć, aby wygenerować odpowiedź. W zależności od usługi i ustawień, Twoje dane mogą również:

- Przechowywane lub rejestrowane (w celu podniesienia jakości usług, rozwiązywania problemów, zapewnienia bezpieczeństwa lub zgodności)
- W niektórych przypadkach poddane przeglądowi (np. w celu zwiększenia bezpieczeństwa lub jakości)
- Użyte do dostosowania Twojego doświadczenia (w niektórych narzędziach)
- Odzwierciedlone w rezultatach
- Udostępniane pośrednio podczas kopiowania i wklejania wyników do innych aplikacji lub przesyłania zrzutów ekranu.

Chcesz wiedzieć, dlaczego ograniczenie udostępniania danych jest istotne?

- Zmniejszasz ryzyko naruszenia prywatności (Twoje dane nie mogą zostać ujawnione, jeśli ich nie udostępniłeś).
- Redukujesz ryzyko wyrządzenia szkody innym (mniej danych osób trzecich, mniej identyfikatorów).
- Zmniejszasz ryzyko wyciągnięcia niesprawiedliwych lub mylących wniosków (mniej „szumu”, więcej informacji skoncentrowanych na celu).
- Posiadasz większą kontrolę nad swoimi danymi.

Chcesz zapoznać się z SAFE Workflow Ladder?

Zasada stopu:

Jeśli nie masz możliwości usunięcia elementów wrażliwych lub poufnych
→ nie przesyłaj!

Jeśli potrzebujesz wsparcia ze strony AI, nie rozpoczynaj od przesyłania wszystkiego. Wybierz najbezpieczniejszy poziom, który nadal funkcjonuje.

- **Poziom 1 (najbezpieczniejszy)**: Zadaj pytanie ogólne (bez informacji osobistych).
  Przykład: „Jak stworzyć CV?”
- **Poziom 2:** Skorzystaj z anonimowego podsumowania (role, bez identyfikatorów).
  Przykład: „Osoba A i Osoba B miały spór dotyczący terminów – w jaki sposób mogę zareagować?”
- **Poziom 3**: Wykorzystuj ograniczone, zanonimizowane dane jedynie w przypadku konieczności.
  Przykład: „Oto zestawienia według kategorii (bez nazw). Proszę o pomoc w interpretacji trendów.”

Jak uczynić niebezpieczne zapytanie bezpiecznym (przed → po)

Zapytanie przed:

„Oto moje kompletne CV. Proszę o jego poprawę.”

Co jest ryzykowne: zbędne identyfikatory, dane kontaktowe osób trzecich, informacje, które można skopiować lub wykraść.

Zapytanie po:

„Proszę o poprawę sformułowania i struktury tego CV. Usunąłem dane osobowe. Proszę skupić się na klarowności, osiągnięciach oraz zwięzłym wyrażeniu.”

Reguła

Udostępniaj wzorzec i cel, a nie informacje osobiste.

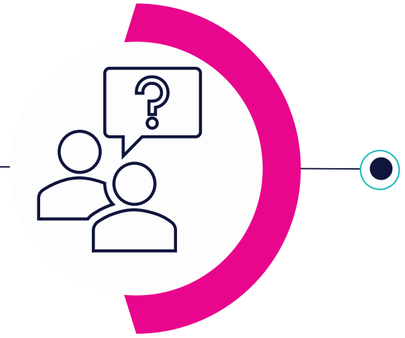
Bezpieczeństwo wyników: wyniki sztucznej inteligencji mogą również prowadzić do wycieku informacji.

Nawet jeśli udostępniasz informacje z ostrożnością, dane wyjściowe mogą wciąż zawierać poufne szczegóły, jeśli:

- Twój wkład zawierał informacje prywatne lub
- Sztuczna inteligencja powtarza to w podsumowaniu, na liście kontrolnej lub w projekcie wiadomości.

Bezpieczniejsze praktyki:

- Traktuj wyniki jako dokumenty: nie umieszczaj ich w publicznych czatach ani grupach.
- Przed udostępnieniem usuń nazwy, identyfikatory oraz unikalne szczegóły.
- Jeśli dane wyjściowe odnoszą się do konkretnej osoby, zachowaj je dla siebie lub przedstaw je w sposób bardziej ogólny.
- Przechowuj poufne dane wyjściowe w bezpiecznym miejscu lub usuń je, gdy przestaną być potrzebne.



Kluczowa aktywność: Przekształć niebezpieczne korzystanie ze sztucznej inteligencji w bezpieczne korzystanie ze sztucznej inteligencji (10 minut)

Będziesz doskonalić całą umiejętność, posługując się fikcyjnymi przykładami. Twoim celem jest formułowanie bardziej bezpiecznych żądań, które wciąż rozwiążą problem.

Dla każdego zapytania:

- Zidentyfikuj niebezpieczne elementy (Nigdy nie udostępniaj / Reguła zgody / Ryzyko integralności)
- Przepisz w sposób bezpieczny (minimalizuj, anonimizuj, streszczaj)
- Dodaj jeden wiersz FDC: „Czego może brakować lub co może być tendencyjne?”
- Wybierz przepływ pracy: Poziom 1 / Poziom 2 / Poziom 3 / Zatrzymaj

Rezultat:

- Krótki „Podręcznik bezpiecznego korzystania z danych”, który możesz wykorzystać w codziennym użytkowaniu sztucznej inteligencji.

Kluczowa aktywność: Przekształć niebezpieczne korzystanie ze sztucznej inteligencji w bezpieczne korzystanie ze sztucznej inteligencji (10 minut)

Ćwiczenie 1:

Zrzut ekranu rozmowy (prywatność + zgoda)

„Oto zrzut ekranu z rozmowy z moim znajomym [widoczne imiona]. Proszę, powiedzcie mi, kto ma rację i jak powinienem odpowiedzieć.”

Zadania etapowe:

- **NSL (Nigdy nie udostępniaj): Czego nigdy nie powinno się tutaj przesyłać?**
- **Reguła zgody: Czy posiadasz zgodę na udostępnianie tych danych?**
- **Bezpieczne przepisywanie (poziom 1 lub 2): Zasięgnij porady bez zrzutu ekranu.**
- **Granica sprawiedliwości: Czego może brakować w tej narracji?**

……. Linia FDC:……

Kluczowa aktywność: Przekształć niebezpieczne korzystanie ze sztucznej inteligencji w bezpieczne korzystanie ze sztucznej inteligencji (10 minut)

Ćwiczenie 2:

Lista osób (poufność + uczciwość)

„Proszę o przeanalizowanie tej listy wolontariuszy [z nazwiskami, numerami telefonów oraz problemami z frekwencją]. Oceń, kto zasługuje na zaufanie.”

Dlaczego to jest ryzykowne:

- Identyfikatory osobiste oraz dane osób trzecich
- Szkodliwy rezultat (klasyfikacja osób)
- Wysokie ryzyko integralności i oddziaływania
- **Twój bezpieczniejszy cel**: **zadaj pytanie dotyczące ulepszenia systemu, a nie oceniania jednostek.**

05

Niektóre informacje

(15 minut)

Kradzież danych i poświadczeń – zauważalny jest rosnący trend nielegalnego wykorzystywania danych przy użyciu sztucznej inteligencji.

„7% wzrost liczby ataków ransomware oraz innych destrukcyjnych incydentów. 23% wzrost liczby prób kradzieży danych uwierzytelniających.

Microsoft Threat Intelligence, „Raport Microsoft Digital Defense 2025: Oświetlanie ścieżki do bezpiecznej przyszłości" (Microsoft, 2025); https://www.microsoft.com/en-us/security/security-insider/threat-landscape/microsoft-digital-defense-report-2025

„Wielkość wycieku danych dla dziesięciu głównych rodzin ransomware wzrosła o 92,7%."

Zscaler ThreatLabz, „Raport Zscaler ThreatLabz 2025 dotyczący oprogramowania ransomware" (Zscaler, 2025); https://threatlabz.zscaler.com/

Kradzież danych i poświadczeń – zauważalny jest rosnący trend nielegalnego wykorzystywania danych przy użyciu sztucznej inteligencji.

„Liczba ataków ransomware na organizacje przemysłowe wzrosła o 87% w roku 2024.”

Dragos, ósmy coroczny przegląd cyberbezpieczeństwa OT jest już dostępny (2025); https://www.dragos.com/blog/dragos-8th-annual-ot-cybersecurity-year-in-review-is-now-available.

Kradzież danych i poświadczeń – zauważalny jest rosnący trend nielegalnego wykorzystywania danych przy użyciu sztucznej inteligencji.

„Podmiot [...] w znacznym stopniu polegał na Claude w zakresie implementacji [złośliwego oprogramowania]”.

A. Moix, K. Lebedev, J. Klein, „Raport o zagrożeniach wywiadowczych: sierpień 2025” (Anthropic, 2025); https://www-cdn.anthropic.com/b2a76c6f6992465c09a6f2fce282f6c0cea8c200.pd

„Operatorzy ransomware APT INC zaimplementowali skrypt do niszczenia danych, prawdopodobnie autorstwa LLM.”

CrowdStrike, „Globalny raport o zagrożeniach CrowdStrike 2025” (CrowdStrike, 2025); https://www.crowdstrike.com/en-gb/global-threat-report/

Kradzież danych i poświadczeń – zauważalny jest rosnący trend nielegalnego wykorzystywania danych przy użyciu sztucznej inteligencji.

„[Google Threat Intelligence Group] zidentyfikowała rodzinę kodów, która wykorzystywała możliwości sztucznej inteligencji podczas wykonywania kodu, aby dynamicznie modyfikować zachowanie złośliwego oprogramowania. […] Atakujący wykraczają poza […] użycie narzędzi sztucznej inteligencji do wsparcia technicznego.”

Google Threat Intelligence Group, „GTIG AI Threat Tracker: Postępy w wykorzystaniu narzędzi AI przez podmioty stwarzające zagrożenie” (Google Threat Intelligence, 2025); https://services.google.com/fh/files/misc/advances-in-threat-actor-usage-of-ai-tools-pl.pdf

Kradzież danych i poświadczeń

Udział systemów sztucznej inteligencji w tym trendzie wydaje się być ograniczony i prawdopodobnie ma drugorzędne znaczenie w porównaniu do innych czynników. Jest jednak możliwe, że niektórzy złośliwi napastnicy nie przeprowadziliby ataków bez wykorzystania systemów opartych na sztucznej inteligencji.

Międzynarodowy raport dotyczący bezpieczeństwa sztucznej inteligencji 2026 (DSIT 2026/001, 2026); https://internationalaisafetyreport.org

Kradzież danych i poświadczeń

Ping Identity. (2024). Redukcja luki zaufania w erze sztucznej inteligencji: badanie konsumenckie z 2025 r. Ping Identity. https://www.pingidentity.com/en/lp/2025-consumer-survey.html

Którym usługom internetowym najmniej ufasz w zakresie ochrony swoich danych osobowych?

Pamiętajmy, że sztuczna inteligencja jest obecnie szeroko wykorzystywana i często odgrywa kluczową rolę w świadczeniu usług oraz realizacji transakcji.

Kradzież danych i poświadczeń

Ping Identity. (2024). Redukcja luki zaufania w erze sztucznej inteligencji: badanie konsumenckie z 2025 r. Ping Identity. https://www.pingidentity.com/en/lp/2025-consumer-survey.html

W jaki sposób aktualnie wykorzystujesz sztuczną inteligencję (AI)?

„Chociaż wiedza stanowi potęgę, rosnące zastosowanie sztucznej inteligencji w codziennym życiu uświadomiło wielu konsumentom jej potencjał. Doświadczają na własnej skórze, jak potężna może być ta technologia, gdy wykorzystują ją w swoich aplikacjach. Zauważają również coraz więcej informacji i danych dotyczących wzrostu zagrożeń bezpieczeństwa oraz cyberataków. W rezultacie zrozumiałe jest, że przyczynia się to do kryzysu zaufania wobec zdolności marek do ochrony danych osobowych konsumentów.”

Kradzież danych i poświadczeń

Ping Identity. (2024). Redukcja luki zaufania w erze sztucznej inteligencji: badanie konsumenckie z 2025 r. Ping Identity. https://www.pingidentity.com/en/lp/2025-consumer-survey.html

W jaki sposób aktualnie wykorzystujesz sztuczną inteligencję (AI)?

Kradzież danych i poświadczeń - jakie masz obawy dotyczące sztucznej inteligencji w kontekście bezpieczeństwa tożsamości, jeśli takie występują?

Naruszenie mojej prywatności przez systemy sztucznej inteligencji.	27%
Phishing generowany przez sztuczną inteligencję.	27%
Sztuczna inteligencja jest wykorzystywana do nieautoryzowanego podszywania się pod mnie lub markę/osobę, z którą nawiązuję interakcję.	26%
Brak przejrzystości dotyczącej metod, w jakie systemy sztucznej inteligencji wykorzystują i przechowują moje dane osobowe.	25%
Zwiększone ryzyko cyberbezpieczeństwa – stanowiące zagrożenie dla bezpieczeństwa moich danych osobowych.	25%
Klonowanie głosu za pomocą sztucznej inteligencji	24%

Ciąg dalszy na kolejnym slajdzie

Ping Identity. (2024). Redukcja luki zaufania w erze sztucznej inteligencji: badanie konsumenckie z 2025 roku. Ping Identity. https://www.pingidentity.com/en/lp/2025-consumer-survey.html

Kradzież danych i poświadczeń - jakie masz obawy dotyczące sztucznej inteligencji w kontekście bezpieczeństwa tożsamości, jeśli takie występują?

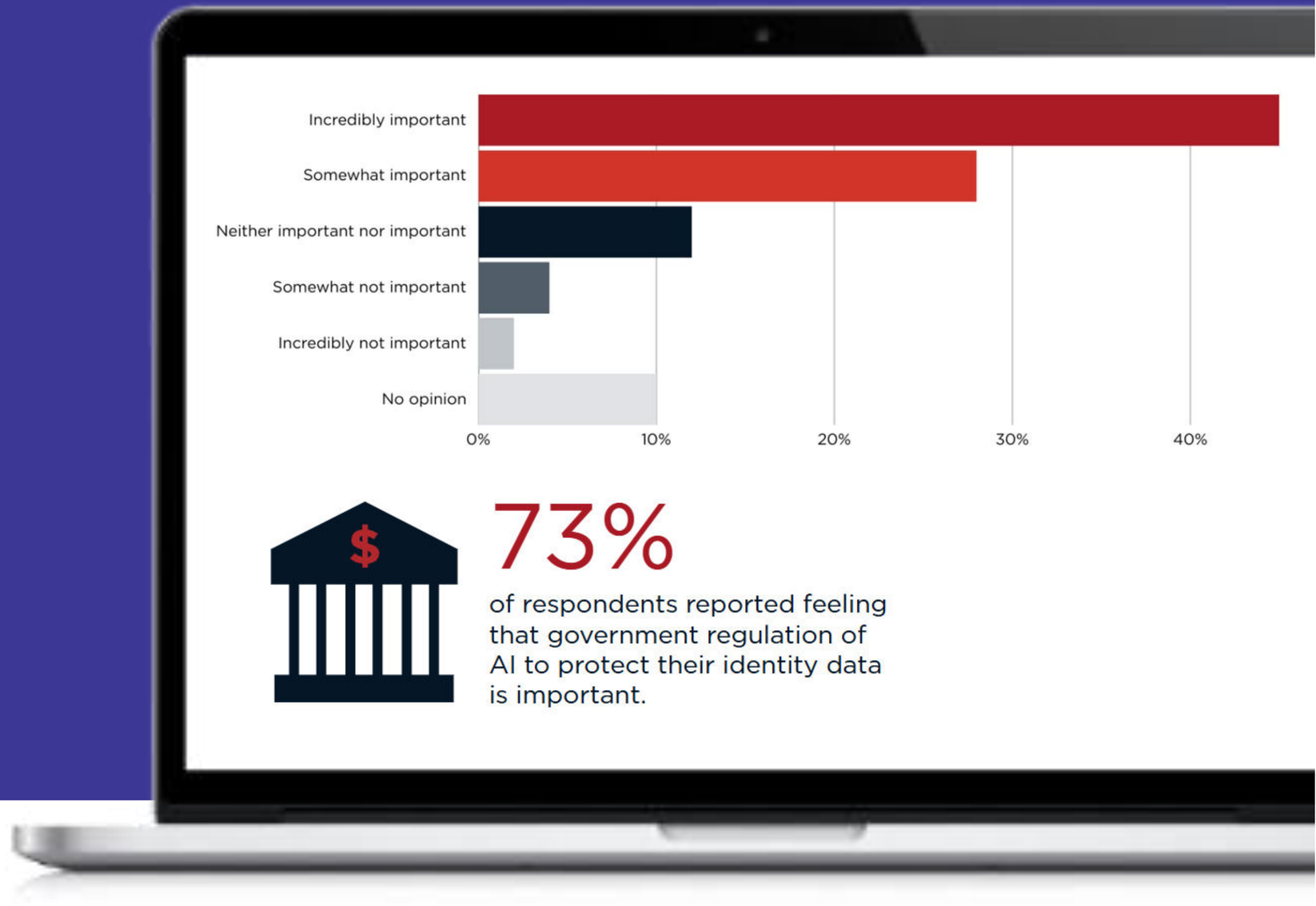
Deepfake’owe odwzorowania osób, które znam i którym ufam.	22%
Podszywanie się pod sztuczną inteligencję za pomocą chatbotów	20%
Brak precyzji/stronniczość w systemach sztucznej inteligencji	18%
Nie rozumiem, jakie mam zabezpieczenia prawne dotyczące sposobu, w jaki sztuczna inteligencja przetwarza moje dane.	18%
Autoryzacja sztucznej inteligencji do działania w moim imieniu	17%
Nie odczuwam żadnych obaw dotyczących sztucznej inteligencji oraz bezpieczeństwa tożsamości.	9%
Nie jestem pewien.	13%

Ping Identity. (2024). Redukcja luki zaufania w erze sztucznej inteligencji: badanie konsumenckie z 2025 roku. Ping Identity. https://www.pingidentity.com/en/lp/2025-consumer-survey.html

Kradzież danych i poświadczeń

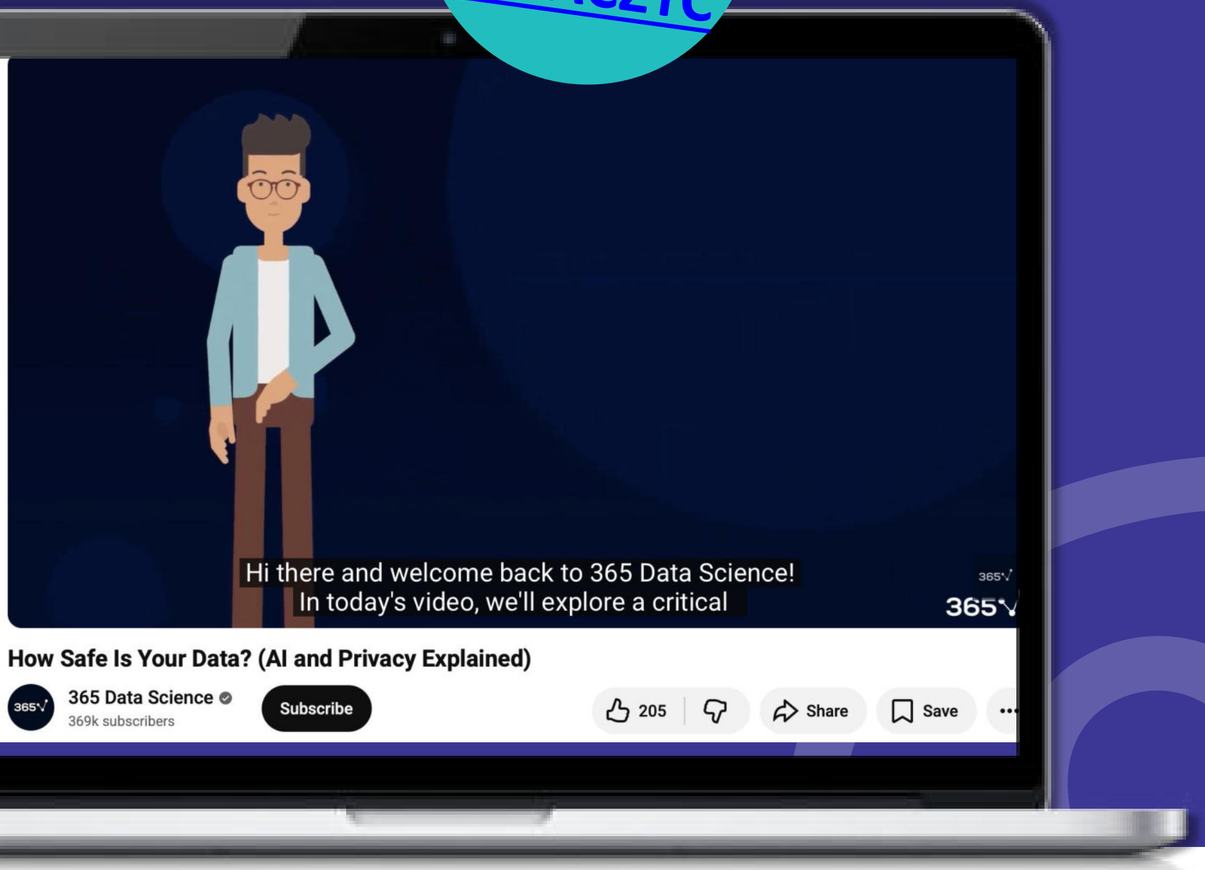
Ping Identity. (2024). Redukcja luki zaufania w erze sztucznej inteligencji: badanie konsumenckie z 2025 r. Ping Identity. https://www.pingidentity.com/en/lp/2025-consumer-survey.html

W jaki sposób aktualnie wykorzystujesz sztuczną inteligencję (AI)?



Kradzież danych i poświadczeń

KLIKNIJ, ABY ZOBACZYĆ



W dzisiejszej erze cyfrowej zabezpieczenie danych jest istotniejsze niż kiedykolwiek.

Film ten ukazuje skomplikowane interakcje między sztuczną inteligencją a danymi, którymi codziennie dzielimy się w sieci – od postów w mediach społecznościowych po nawyki zakupowe online.

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.

AI Skills, Powered by Values

values

www.valuesai.eu

MODUŁ 06

Sztuczna inteligencja oraz ślad cyfrowy

Kontrola tego, co pozostawiasz za sobą

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

01	Cyfrowy ślad (10 min)	3
02	Rodzaje śladów (15 min)	8
03	„Efekt domina”: konsekwencje dla Ciebie i innych (25 min)	15
04	Mój wybór: zwięzła lista kontrolna + mój plan działania (15 min)	25
05	Jak zabezpieczyć swój cyfrowy ślad (10 min)	35
06	Audyt, quiz oraz test (20 min)	39

Materiał dostępny na licencji Creative Commons CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

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

Cyfrowy ślad

(10 minut)

Czym jest cyfrowy ślad?

Cyfrowy ślad to zbiór informacji dotyczących osoby lub organizacji, które pozostają w sieci w wyniku ich aktywności online. Obejmuje on profile w mediach społecznościowych, posty, komentarze, historię przeglądania, zapytania wyszukiwania oraz dane automatycznie rejestrowane przez strony internetowe i aplikacje (np. pliki cookie, adres IP, dane lokalizacyjne).

AI Skills,
Powered
by Values
values

Dlaczego Twój ślad cyfrowy ma znaczenie?

Twój ślad cyfrowy może wpływać na:

1

Prywatność

kto i co wie o tobie

2

Reputację

jak jesteś postrzegany

3

Bezpieczeństwo

co może być użyte w niewłaściwy sposób

4

Inne osoby

Twoje posty mogą również ujawnić innych

Dlaczego Twój cyfrowy ślad jest istotny dla sztucznej inteligencji

Zadaj sobie pytanie:
Gdzie dostrzegam „decyzje algorytmiczne” w aplikacjach, których codziennie używam?

Sztuczna inteligencja integruje się z Twoim cyfrowym śladem, ponieważ:

- Sztuczna inteligencja często wykorzystuje dane behawioralne (kliknięcia, czas oglądania, polubienia, wyszukiwania) do przewidywania, czego będziesz pragnąć następnym razem.
- Twój bierny ślad (uprawnienia, sygnały lokalizacji, dane urządzenia) może wpływać na to, jakie wnioski systemy wyciągną na Twój temat.
- Sztuczna inteligencja ma zdolność wpływania na widoczność i zasięg (jakie treści są promowane, rekomendowane lub ukrywane).
- Twoje decyzje wpływają również na informacje dotyczące innych osób (tagowanie, udostępnianie zdjęć, zrzutów ekranu).

Wykorzystanie sztucznej inteligencji w zgodzie z wartościami

Pamiętaj:
Etyczna sztuczna inteligencja rozpoczyna się od codziennych decyzji cyfrowych.

Cztery pytania dotyczące wartości, którymi warto się kierować przy podejmowaniu decyzji cyfrowych (szczególnie w kontekście sztucznej inteligencji):

- **Decyzyjność**: Czy podejmuję świadomy wybór, czy też klikam automatycznie?
- **Prywatność**: Czy udostępniam więcej informacji (o sobie lub innych), niż jest to konieczne?
- **Sprawiedliwość**: Czy może to wyrządzić komuś szkodę lub zaszkodzić (teraz lub w przyszłości)?
- **Przejrzystość**: Czy jestem świadomy, kto ma dostęp do tych informacji i w jaki sposób mogą być one udostępniane?

02

Rodzaje śladów

(15 minut)

Aktywny ślad w sieci

Aktywny ślad cyfrowy to informacje, które świadomie generujesz podczas korzystania z internetu. Przykłady:

- Publikowanie informacji, komentarzy, zdjęć,
- Udostępnianie danych w formularzach i ankietach,
- Tworzenie oraz uzupełnianie profili na stronach internetowych, w przedsiębiorstwach lub na platformach.

Przy tym typie śladu posiadasz stosunkowo większą kontrolę nad tym, co udostępniasz.

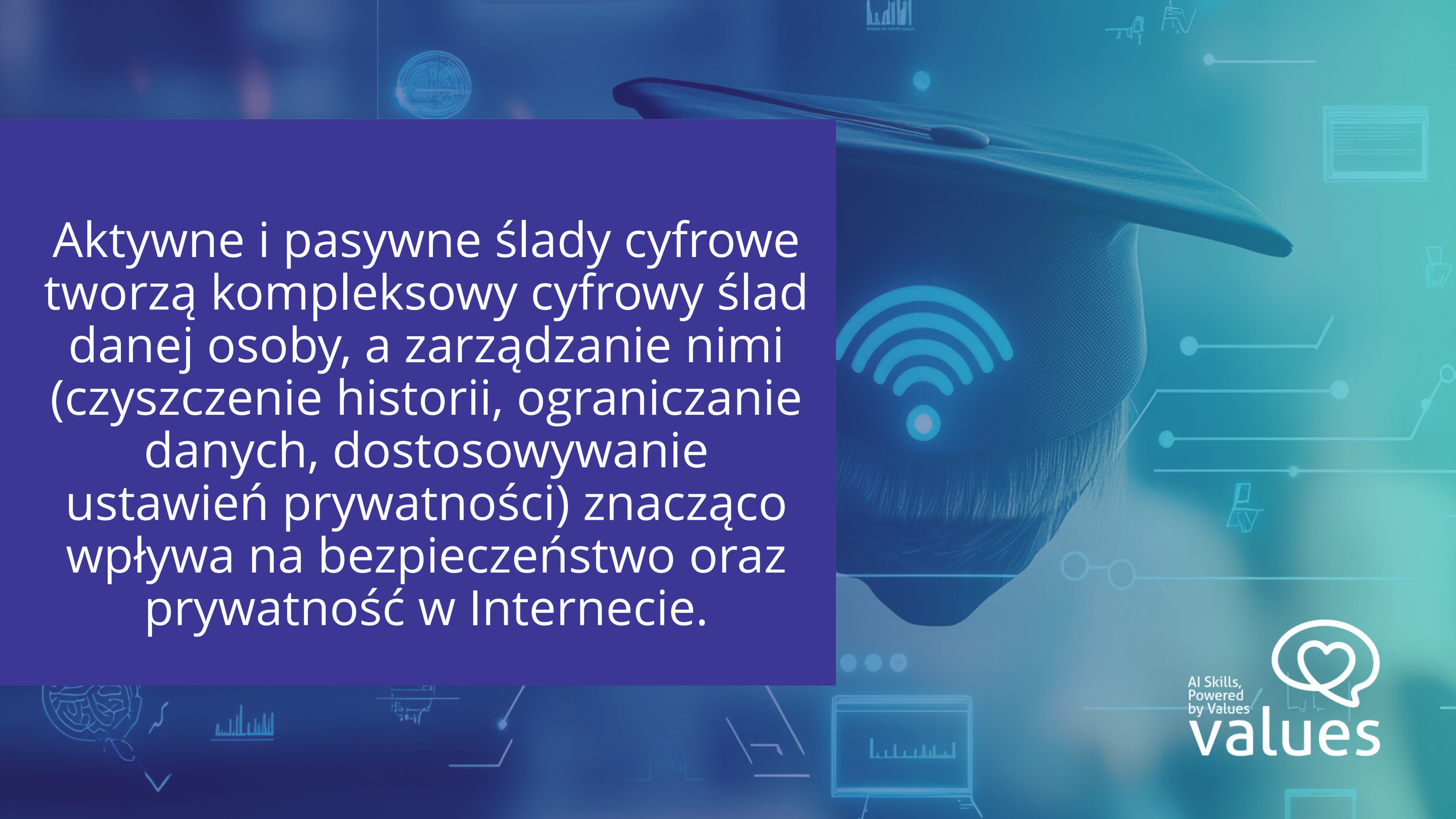
Pasywny ślad cyfrowy

Pasywny ślad cyfrowy to dane gromadzone automatycznie, często bez pełnej świadomości użytkownika. Przykłady:

- Historia przeglądania oraz logi serwera,
- Informacje gromadzone przez pliki cookie oraz moduły śledzące,
- Dane dotyczące urządzenia, systemu oraz lokalizacji (np. adres IP),
- Interakcje rejestrowane przez platformy (kliknięcia, czas trwania wizyty, sesje).

Ślad ten jest zazwyczaj zbierany przez systemy analityczne, reklamowe i śledzące, co utrudnia użytkownikom jego usunięcie lub pełną kontrolę.

Aktywne i pasywne ślady cyfrowe tworzą kompleksowy cyfrowy ślad danej osoby, a zarządzanie nimi (czyszczenie historii, ograniczanie danych, dostosowywanie ustawień prywatności) znacząco wpływa na bezpieczeństwo oraz prywatność w Internecie.



AI Skills,
Powered
by Values
values

Szybka aktywność: Uporządkuj ślady

Zadanie (2 minuty) i wybierz jedną opcję:

Aktywny

Pasywny

Obydwa?

- Oznaczanie znajomej osoby na zdjęciu
- Aplikacja prosi o dostęp do lokalizacji „zawsze”
- Polubienie postu publicznego
- Model Twojego urządzenia w metadanych fotografii
- Publikowanie opowieści z etykietą lokalizacji
- Pliki cookie śledzą Twoje kliknięcia.
- Zrzut ekranu z prywatnego czatu
- Platforma rekomendująca „osoby, które możesz znać”

Podsumowanie: Większość ludzi nie dostrzega swojego „pasywnego” śladu cyfrowego.

Rozważając swój cyfrowy ślad, prawdopodobnie koncentrujesz się głównie na postach, zdjęciach i komentarzach. Niemniej jednak znaczna część Twojego śladu powstaje bez Twojego aktywnego publikowania czegokolwiek, poprzez Twoje ustawienia, sygnały z urządzenia, uprawnienia oraz aktywność w sieci.

AI Skills,
Powered
by Values
values

Gdzie sztuczna inteligencja łączy się z Twoim śladem: codzienne przykłady

Nie musisz „tworzyć ani aktywnie korzystać ze sztucznej inteligencji”, aby na nią wpływać. Wiele usług internetowych wykorzystuje sztuczną inteligencję do personalizacji treści, klasyfikacji informacji oraz automatyzacji decyzji – często opierając się na sygnałach lokalizacyjnych.

Typowe sytuacje z wykorzystaniem sztucznej inteligencji, oparte na sygnałach śladowych:

- Ranking: co pojawia się jako pierwsze w wynikach wyszukiwania lub w Twoim strumieniu
- Targetowanie: reklamy spersonalizowane i „sugerowane treści”
- Filtrowanie i moderacja: treść oznaczona, ograniczona w zasięgu lub usunięta
- Wykrywanie wzorców: nietypowe powiadomienia logowania lub monity o „podejrzanej aktywności”
- Rekomendacje: kanały „Dla Ciebie”, proponowane filmy, muzyka lub produkty

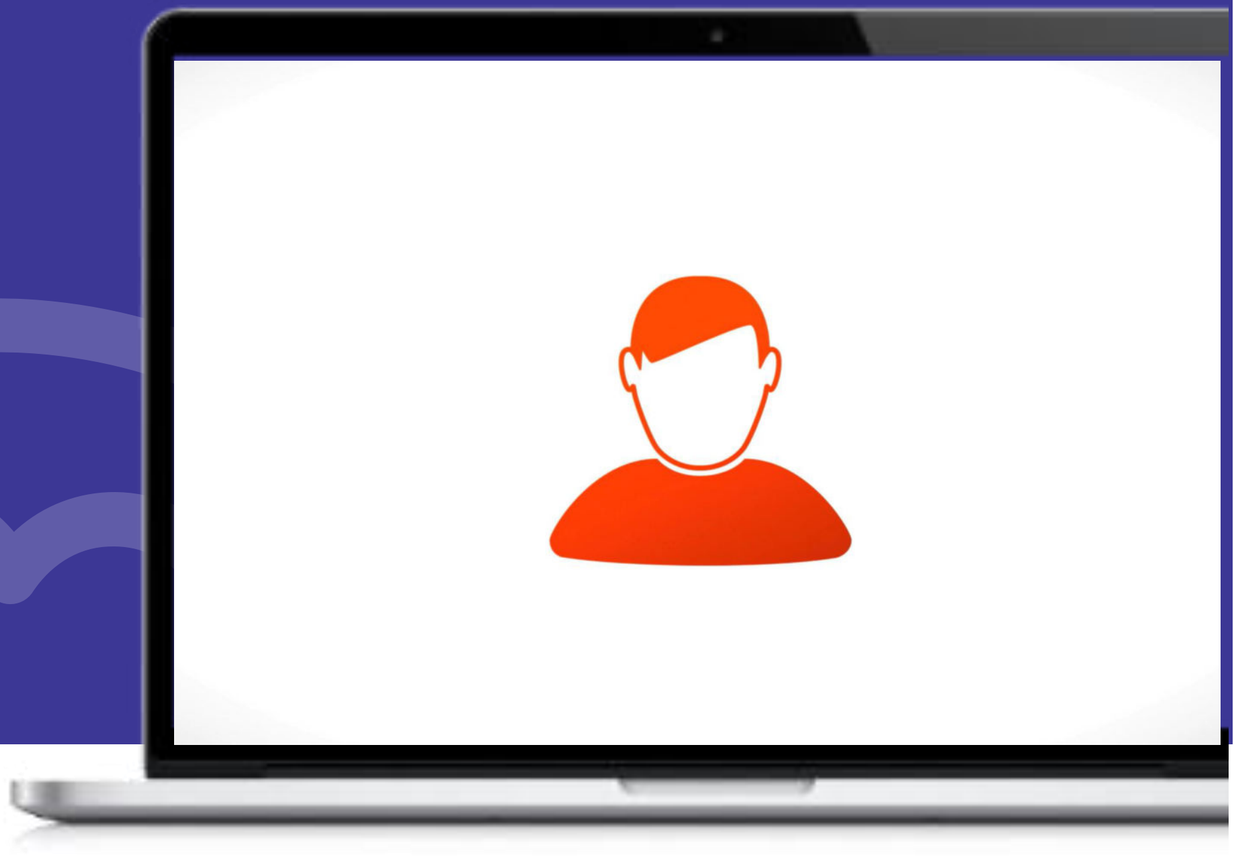
03

„Efekt domina”: konsekwencje dla Ciebie i innych

(25 minut)

„Efekt domina”: działania w sieci mają szerszy zasięg, niż się spodziewasz.

Treści online mogą szybko przemieszczać się między ludźmi i platformami. Nawet drobne działania (komentarz, tag, zrzut ekranu) mogą się rozprzestrzeniać, tracić kontekst i pozostawać dostępne dłużej, niż pierwotnie zamierzono.



„Efekt domina”: działania w sieci mają szerszy zasięg, niż się spodziewasz.

Jeden post może prowadzić do:

- **Kopie**: zrzuty ekranu, udostępnianie, pobieranie
- **Utrata kontekstu**: żarty lub emocje są niewłaściwie interpretowane bez wyjaśnienia.
- **Długa żywotność**: przestarzałe treści mogą się pojawić ponownie po miesiącach lub latach.
- **Szersza publiczność**: znajomi znajomych, grupy lub publiczne udostępnianie
- **Wpływ na innych**: osoby obecne w Twoich treściach mogą również zostać dotknięte (prawdopodobnie bez ich zgody).

Konsekwencje dla Ciebie: prywatność, wizerunek, bezpieczeństwo oraz samopoczucie

Cyfrowy ślad nie jest automatycznie „dobry” ani „zły”. Jego ocena zależy od kontekstu oraz dokonanych wyborów. Niemniej jednak, pewne wzorce (nadmierne udostępnianie, otwarte ustawienia, impulsywne publikowanie) mogą stwarzać zagrożenia dla Twojej prywatności, reputacji, a nawet bezpieczeństwa, których można uniknąć.

Prywatność (kto może się czegoś o tobie dowiedzieć)	**Reputacja** (jak jesteś postrzegany)	**Bezpieczeństwo** (co może być niewłaściwie wykorzystane)	**Dobre samopoczucie.** (stres i konflikt)
Publiczne profile, tagi, lokalizacja oraz uprawnienia mogą ujawniać więcej, niż pierwotnie zamierzano.	Posty można czytać bez kontekstu; pierwsze wrażenie może szybko się uformować i pozostać trwałe.	Nadmierne udostępnianie danych może prowadzić do oszustw, przejęć kont oraz niepożądanych kontaktów (nie tylko w kontekście życia cyfrowego).	Kłótnie, nękanie lub poczucie żalu po opublikowaniu posta mogą negatywnie wpływać na pewność siebie oraz nastrój.

Konsekwencje dla CIEBIE: prywatność, wizerunek, bezpieczeństwo oraz samopoczucie.

Przykład:

Zdjęcie z widoczną lokalizacją oraz otwartymi ustawieniami profilu może ujawnić, gdzie się znajdujesz, z kim spędzasz czas i co robisz na co dzień – bez konieczności bezpośredniego podawania informacji.

Zadaj sobie pytanie: Które ryzyko jest dla mnie obecnie najistotniejsze: prywatność, reputacja, Bezpieczeństwo czy komfort psychiczny? Dlaczego?

Niewielkie modyfikacje (ustawienia odbiorców, uprawnienia, wstrzymanie publikacji) mogą znacząco zredukować te ryzyka.

Konsekwencje dla INNYCH: szacunek, akceptacja i rzetelność

Typowe metody, w jakie narażamy innych na niebezpieczeństwo w Internecie:

- Publikowanie lub udostępnianie treści w lokalizacjach, w których obecni są inni użytkownicy.
- Oznaczanie lub wspominanie kogoś (umieszczanie odnośników do treści)
- Udostępnianie zrzutów ekranu z prywatnych rozmów
- Ujawnianie czyjejś lokalizacji, harmonogramu lub danych osobowych.
- Udostępnianie treści, które mogą później prowadzić do konfliktów

Konsekwencje dla INNYCH: szacunek, akceptacja i rzetelność

Przykład

„Zabawny” zrzut ekranu z prywatnego czatu może zostać udostępniony, błędnie zinterpretowany i użyty przeciwko komuś – ponieważ pierwotny kontekst zanika.

Jak kontrolować

Zadaj sobie pytanie: Czy byłoby dla mnie akceptowalne, gdyby ktoś bez pytania udostępnił podobne treści na mój temat?

Ochrona prywatności obejmuje także prywatność innych osób:
Zgoda i szacunek stanowią integralne elementy odpowiedzialnego działania w Internecie.

Mini scenariusze: wybierz najbardziej odpowiednią opcję

Decyzje podejmowane w rzeczywistości często są podejmowane „na szybko”. To ćwiczenie pomoże Ci przećwiczyć wybór opcji, która chroni prywatność, minimalizuje szkody i szanuje innych ludzi, zwłaszcza gdy treść może być kopiowana lub udostępniana.

Scenariusz: Prywatny zrzut ekranu

Znajomy przesyła Ci zrzut ekranu z prywatnej konwersacji i stwierdza: „To jest niezwykle zabawne – podziel się tym z innymi”.

Opcja 1: Wyślij to natychmiast (to zabawne).

Opcja 2: Zachowaj prywatność i zapytaj znajomego, jakie cele ma na myśli, udostępniając tę wiadomość.

Opcja 3: Nie udostępniaj informacji, chyba że osoby zainteresowane wyrażą na to zgodę; w przypadku wątpliwości – nie udostępniaj.

Mini scenariusze: wybierz najbardziej odpowiednią opcję

Decyzje podejmowane w rzeczywistości często są podejmowane „na szybko”. To ćwiczenie pomoże Ci przećwiczyć wybór opcji, która chroni prywatność, minimalizuje szkody i szanuje innych ludzi, zwłaszcza gdy treść może być kopiowana lub udostępniana.

Scenariusz: Prywatny zrzut ekranu

Znajomy przesyła Ci zrzut ekranu z prywatnej konwersacji i stwierdza: „To jest zabawne – podziel się tym z innymi”.

Twoje zadanie (30–60 sekund): Wybierz 1 / 2 / 3 Napisz jedno zdanie:

„Wybrałem... ponieważ....”

Sztuczna inteligencja a prywatność – przykład

O czym mówi przykład:

- Systemy sztucznej inteligencji analizują dane pochodzące z codziennych aktywności (np. mediów społecznościowych, przeglądania stron internetowych, zakupów online).
- Dane osobowe są wykorzystywane jako dane wejściowe do szkolenia i doskonalenia modeli sztucznej inteligencji.
- Nawet podstawowe zachowania w sieci mogą być zintegrowane w szczegółowe profile użytkowników.

04

Mój wybór: zwięzła lista kontrolna + mój plan działania

(15 minut)

Moje narzędzie selekcji: ZATRZYMAJ SIĘ-POMYŚL-SPRAWDŹ-WYBIERZ

Ta lista kontrolna to prosty nawyk, który pomoże Ci podejmować bezpieczniejsze i bardziej sprawiedliwe decyzje w Internecie. Skorzystaj z niej przed opublikowaniem, udostępnieniem, oznaczeniem, przesłaniem zrzutu ekranu lub zaakceptowaniem uprawnień aplikacji, zwłaszcza gdy czujesz się pośpieszony lub zdenerwowany.

ZATRZYMAJ SIĘ	POMYŚL	SPRAWDŹ	WYBIERZ
(zatrzymaj się na 5 sekund)	**(cel i odbiorcy):**	**(prywatność + ludzie + informacje)**	**(najlepszy kolejny krok)**
Czy publikuję szybko, znużony, zdenerwowany, podekscytowany lub pod presją?	Jaki jest cel (informowanie, żartowanie, popisywanie się, prośba o pomoc)? Kto obecnie to zobaczy, a kto może to zobaczyć w przyszłości?	Czy ustawienia odbiorców są poprawne (publiczne / znajomi / bliscy znajomi / grupa)? Czy obejmują inne osoby (twarze, imiona, wiadomości, lokalizacje)? Czy ujawniają dodatkowe informacje (naklejkę z lokalizacją, szczegóły tła, metadane)?	Publikuj / edytuj / przycinaj / rozmywaj / ograniczaj odbiorców / pytaj o zgodę / nie udostępniaj.

Scenariusz 1. Ćwiczenie: wykorzystaj listę kontrolną przed publikacją

Tworzysz pasywny cyfrowy ślad nie tylko podczas przeglądania stron, ale również poprzez to, na co zezwalasz aplikacjom.

Ten scenariusz pozwoli Ci ćwiczyć w zakresie wyboru niezbędnych uprawnień oraz odrzucania tych, które gromadzą więcej danych, niż jest to konieczne.

Zobacz kolejny slajd

Scenariusz 1:

Instalujesz nową aplikację, taką jak edytor zdjęć, gra lub aplikacja edukacyjna.

Podczas konfiguracji prosi o:

- Lokalizacja
- Kontakty
- Mikrofon
- Zdjęcia

ZATRZYMAJ SIĘ

Czy klikając „Zezwól”, automatycznie przyspieszam zakończenie konfiguracji?

POMYŚL

Czego tak naprawdę potrzebuje aplikacja, aby realizować swoją podstawową funkcję? Czy mogę używać aplikacji z ograniczoną liczbą uprawnień?

SPRAWDŹ

Lokalizacja: czy lokalizacja jest niezbędna „zawsze”, „tylko podczas używania”, czy wcale? Kontakty: czy dostęp do nich jest rzeczywiście konieczny, czy jedynie „pożądany”? Mikrofon: czy jest wymagany do funkcji, z której zamierzam korzystać? Zdjęcia/Media: czy mogę wybrać „tylko wybrane zdjęcia” zamiast „wszystkich zdjęć”?

WYBIERZ

Udziel minimalnych wymaganych uprawnień. Preferuj opcję „tylko podczas używania” zamiast „zawsze”. Przejrzyj i dostosuj uprawnienia później, jeśli funkcjonalność rzeczywiście ich wymaga.

Scenariusz 2. Ćwiczenie: zastosuj listę kontrolną, gdy zaangażowane są inne osoby

Jeżeli w Twoich treściach uczestniczą inne osoby, Twój wybór wpłynie również na nie.

Ten scenariusz umożliwi Ci ćwiczenie podejmowania decyzji z uwzględnieniem prywatności i zgody, ponieważ działania związane z udostępnianiem, tagowaniem i przekazywaniem mogą być trudne do odwrócenia.

Zobacz kolejny slajd

Scenariusz 2:

Masz grupowe zdjęcie z spotkania, wydarzenia lub wieczornego wyjścia. Chcesz je opublikować i oznaczyć znajomych. Jedna osoba na zdjęciu wcześniej zadeklarowała, że nie życzy sobie publikacji w internecie.

ZATRZYMAJ SIĘ

Czy zamierzam opublikować coś szybko, ponieważ pragnę podzielić się tą chwilą już teraz?

POMYŚL

Jaki jest cel publikacji? Czy tagowanie jest niezbędne? Kto zobaczy to teraz – a kto będzie mógł zobaczyć później?

SPRAWDŹ

Zgoda: Czy wszyscy wyrazili zgodę na wyświetlanie lub oznaczanie / Publiczność: Czy jest to dostępne publicznie, czy ograniczone do wąskiej grupy / Ryzyko: Czy to zdjęcie może w przyszłości kogoś skrzywdzić lub zawstydzić? Alternatywy: Czy mogę przyciąć, rozmyć lub wybrać inne zdjęcie?

WYBIERZ

Wybierz optymalny następny krok: najpierw zapytaj daną osobę / opublikuj, nie ujawniając jej treści / rozmyj lub przytnij / ogranicz liczbę odbiorców / nie publikuj.

Scenariusz 3. Ćwiczenie: wykorzystaj listę kontrolną uprawnień aplikacji

Tworzysz pasywny cyfrowy ślad nie tylko podczas przeglądania stron, ale również poprzez to, na co zezwalasz aplikacjom.

Ten scenariusz umożliwi Ci praktykę w zakresie wyboru niezbędnych uprawnień oraz odrzucania tych, które gromadzą więcej danych, niż jest to konieczne.

Zobacz kolejny slajd

Scenariusz 3:

Instalujesz nową aplikację, taką jak edytor zdjęć, gra lub aplikacja edukacyjna.

Podczas konfiguracji prosi o:

- Lokalizacja
- Kontakty
- Mikrofon
- Zdjęcia

ZATRZYMAJ SIĘ

Czy klikając „Zezwól”, automatycznie przyspieszam zakończenie konfiguracji?

POMYŚL

Czego tak naprawdę potrzebuje aplikacja, aby realizować swoją podstawową funkcję? Czy mogę używać aplikacji z ograniczoną liczbą uprawnień?

SPRAWDŹ

Lokalizacja: czy jest ona niezbędna „zawsze”, „tylko w trakcie użytkowania” czy w ogóle? **Kontakty**: czy dostęp do nich jest rzeczywiście niezbędny, czy też jedynie „pożądany”? **Mikrofon**: czy jest niezbędny do funkcji, z której zamierzam korzystać? **Zdjęcia/Media**: czy mogę wybrać „tylko wybrane zdjęcia” zamiast „wszystkich zdjęć”?

WYBIERZ

Bezpieczniejsze podejście domyślne: Przyznawaj jedynie minimalne wymagane uprawnienia. Preferuj opcję „tylko podczas używania” zamiast „zawsze”. Przejrzyj i dostosuj uprawnienia później, jeśli funkcjonalność rzeczywiście ich wymaga.

Trzy elementy o znaczącym wpływie: niewielkie zmiany, ogromny efekt

Nie musisz znać wszystkich ustawień prywatności, aby poprawić swój cyfrowy ślad. Te trzy elementy sterujące znacząco wpływają na większość platform i aplikacji – warto je zatem sprawdzić, nawet jeśli dysponujesz ograniczonym czasem.

Twój cyfrowy ślad rośnie najszybciej, gdy zarządzasz odbiorcami, lokalizacją i widocznością.

Trzy elementy o znaczącym wpływie: niewielkie zmiany, ogromny efekt

Podstawowa treść (ogólny przewodnik do samodzielnej nauki)

- **Widoczność odbiorców** (kto ma dostęp do Twoich treści): Upewnij się, czy domyślnie jest to grupa publiczna, znajomi, bliscy znajomi czy prywatna. Pamiętaj: „znajomi znajomych” lub grupy otwarte mogą nadal obejmować szeroką publiczność.

- **Udostępnianie lokalizacji** (gdy Twoja lokalizacja jest widoczna): Unikaj publicznego udostępniania lokalizacji w czasie rzeczywistym. Preferuj dostęp do lokalizacji „tylko podczas korzystania” (lub wyłącz go, jeśli nie jest potrzebny). Zwracaj uwagę na ukryte wskazówki dotyczące lokalizacji na zdjęciach (w tle, w punktach orientacyjnych, popularnych miejscach).

- **Możliwość odnalezienia i synchronizacja kontaktów** (jak inni mogą Cię znaleźć): Zweryfikuj, czy Twój profil jest dostępny za pośrednictwem numeru telefonu lub adresu e-mail. Sprawdź czy aplikacje domyślnie synchronizują Twoje kontakty. Zdecyduj, czy pragniesz być wyszukiwany, czy też aby kontaktowały się z Tobą jedynie wybrane osoby.

05

Jak zabezpieczyć swój cyfrowy ślad

(10 minut)

Praktyczne zasady.

01 Ograniczaj udostępnianie informacji w mediach społecznościowych.

02 Zaostrzenie ustawień prywatności w mediach społecznościowych

03 Ogranicz ilość danych przesyłanych do sieci.

05 Od czasu do czasu usuwaj pliki cookie przeglądarki oraz inne pliki śledzące z urządzeń osobistych.

Praktyczne zasady

06 Korzystaj z usług VPN, gdy to możliwe.

07 Anonimizuj, nie zezwalaj na dostęp do lokalizacji śledzenia przez aplikacje lub go ograniczaj.

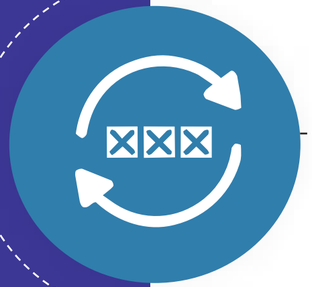
08 Zainstaluj oraz zaktualizuj oprogramowanie antywirusowe na urządzeniach osobistych.

09 Okresowo weryfikuj informacje finansowe/kredytowe oraz medyczne/zdrowotne.

Praktyczne zasady

10 Nie otwieraj załączników ani nie klikaj w linki pochodzące z nieznanych lub wątpliwych źródeł.

11 Skonfiguruj powiadomienia o raportach kredytowych.



12 Ciągle aktualizuj hasła oraz zabezpieczenia haseł, szczególnie po uzyskaniu informacji o naruszeniu danych.

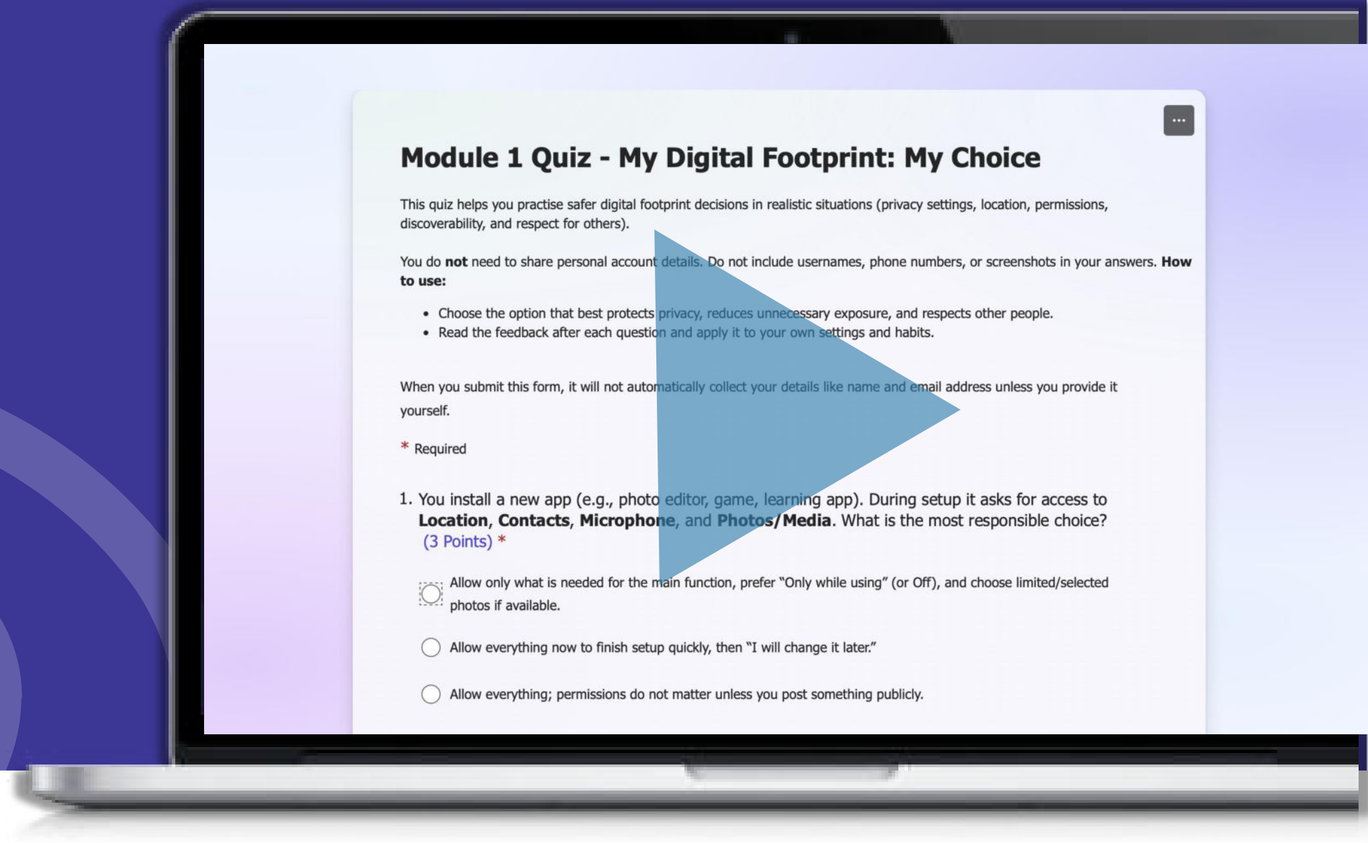
06

Audyt, quiz oraz test

(20 minut)

Quiz

Ten quiz pomoże Ci doskonalić umiejętność podejmowania bezpieczniejszych decyzji dotyczących cyfrowego śladu w realistycznych sytuacjach (ustawienia prywatności, lokalizacja, uprawnienia, wykrywalność oraz poszanowanie innych).



Test: Zbadaj swój cyfrowy ślad

Nie jesteś pewien swojego cyfrowego śladu? Odpowiedz „Tak” lub „Nie” na poniższe pytania, a następnie zsumuj łączną liczbę odpowiedzi „Tak”, aby ocenić złożoność swojego śladu.

Źródło: Cyfrowy ślad osoby wewnątrz organizacji oraz związane z nim ryzyko. Wsparcie w poszukiwaniu pracy. Agencja Kontrwywiadu i Bezpieczeństwa Obrony (CDSE), 2022
https://www.cdse.edu/Portals/124/Documents/jobaids/insider/insiders-digital-footprint.pdf

T N

1. Czy dokonujesz zakupów online?
2. Czy zarejestrowałeś się na kupony, tworząc konto?
3. Czy zarejestrowałeś się lub subskrybowałeś newslettery lub aktualizacje bloga?
4. Czy pobrałeś i korzystałeś z aplikacji do zakupów?
5. Czy założyłeś nowe konto karty kredytowej?
6. Czy używasz aplikacji bankowości mobilnej?
7. Czy nabyłeś lub zbyłeś akcje?
8. Czy kiedykolwiek zarejestrowałeś swój adres e-mail w klubie fitness?
9. Czy korzystasz z opieki zdrowotnej?

Test: Zbadaj swój cyfrowy ślad

Nie jesteś pewien swojego cyfrowego śladu? Odpowiedz „Tak” lub „Nie” na poniższe pytania, a następnie zsumuj łączną liczbę odpowiedzi „Tak”, aby ocenić złożoność swojego śladu.

T N

10. Czy kiedykolwiek korzystałeś z aplikacji do monitorowania swojej aktywności i treningów?
11. Czy subskrybujesz czasopismo internetowe lub źródło informacji?
12. Czy kiedykolwiek zdarzyło Ci się ponownie publikować przeczytane artykuły i informacje?
13. Czy korzystasz z mediów społecznościowych na komputerze lub na urządzeniach mobilnych?
14. Czy nawiązujesz kontakt ze znajomymi w internecie?
15. Czy kiedykolwiek dzieliłeś się informacjami, danymi oraz zdjęciami ze swoimi znajomymi w internecie?
16. Czy kiedykolwiek zarejestrowałeś się w serwisie lub aplikacji randkowej?

Źródło: Cyfrowy ślad osoby wewnątrz organizacji oraz związane z nim ryzyko. Wsparcie w poszukiwaniu pracy. Agencja Kontrwywiadu i Bezpieczeństwa Obrony (CDSE), 2022
https://www.cdse.edu/Portals/124/Documents/jobaids/insider/insiders-digital-footprint.pdf

•Test: Zbadaj swój cyfrowy ślad

WYNIKI

Mniej niż 5 odpowiedzi TAK

Cyfrowy ślad, który jest prostszy niż u większości użytkowników Internetu.

5 – 12 odpowiedzi TAK

Cyfrowy ślad, który jest typowy dla większości użytkowników Internetu.

Ponad 12 odpowiedzi TAK

Cyfrowy ślad, który jest bardziej skomplikowany niż większość użytkowników Internetu.

Źródło: Cyfrowy ślad osoby wewnątrz organizacji oraz związane z nim ryzyko. Wsparcie w poszukiwaniu pracy. Agencja Kontrwywiadu i Bezpieczeństwa Obrony (CDSE), 2022
https://www.cdse.edu/Portals/124/Documents/jobaids/insider/insiders-digital-footprint.pdf

Audyt śladu cyfrowego

Teraz będziesz ćwiczyć identyfikację typowych zagrożeń związanych z odciskiem oraz wybieranie praktycznych usprawnień. Możesz to również zrealizować bez ujawniania danych osobowych, korzystając z fikcyjnego przykładu lub opcjonalnie, sprawdzając jedno ustawienie na swoim urządzeniu w trybie prywatnym.

Prywatne sprawdzenie własnego urządzenia – zweryfikuj jedno ustawienie prywatności lub uprawnień na swoim urządzeniu lub koncie. Jeśli pracujesz w zespole, wykonaj to bez udostępniania ekranu ani danych.

Audyt śladu cyfrowego – wybierz jedną z aplikacji mediów społecznościowych, które posiadasz.

Celem nie jest eliminacja Twojej obecności w sieci, lecz ograniczenie zbędnej ekspozycji oraz świadome zaznaczenie swojego śladu.

Uzupełnij wiersze:

Ryzyko 1 → Rozwiązanie 1: _____ → _____

Ryzyko 2 → Rozwiązanie 2: _____ → _____

Ryzyko 3 → Rozwiązanie 3: _____ → _____

Pierwsze działanie: _____

Krok 1: Zidentyfikuj trzy ryzyka związane z odciskiem

- Zbyt szeroka widoczność publiczna (posty/profile dostępne do odkrycia)
- Ujawniono lokalizację (miejsca w czasie rzeczywistym lub w ramach rutynowych działań)
- Tagowanie/wzmianki bez zgody użytkownika
- Zbyt szerokie uprawnienia (kontakty/mikrofon/lokalizacja „zawsze”)
- Widoczne dane osobowe (szkoła/praca, telefon, adres e-mail, rutyna)

Krok 2: Wybierz 3 usprawnienia (dopasuj jedno do każdego ryzyka)

- Ogranicz odbiorców / przełącz na tryb prywatny / sprawdź domyślną widoczność
- Usuń lub ogranicz udostępnianie lokalizacji.
- Zapytaj o zgodę przed oznaczeniem lub publikowaniem innych osób.
- Zredukuj uprawnienia do „minimum koniecznego”
- Usuń dane poufne z profilu

Krok 3: Określ, jakie będzie Twoje „pierwsze działanie”

- Jaką jedną zmianę wprowadziłbyś jako pierwszą, gdybyś dysponował jedynie 2 minutami?

Audyt śladu cyfrowego

Cyfrowy ślad poprawia się dzięki niewielkim, spójnym działaniom – nie dzięki jednemu idealnemu rozwiązaniu. Ten zwięzły plan pomoże Ci przekształcić dzisiejszy audyt w konkretne kroki, które możesz rzeczywiście zrealizować.

Uczyń to wykonalnym: Jeśli Twój plan wydaje się zbyt rozbudowany, ogranicz go do jednego celu i jednego nawyku. Twój cyfrowy ślad kształtuje się poprzez nawyki. Mały plan, którego się trzymasz, jest lepszy niż duży plan, o którym zapominasz.

1.Jedna zmiana, którą wprowadzę dzisiaj (5 minut)

Przykłady:

- Ogranicz widoczność profilu.
- Usuń udostępnianie lokalizacji.
- Zredukuj jedno uprawnienie.
- Sprawdź zdolność wykrywania

Napisz:
Dzisiaj zamierzam:
……
……
……

Audyt śladu cyfrowego

Cyfrowy ślad poprawia się dzięki niewielkim, spójnym działaniom – nie dzięki jednemu idealnemu rozwiązaniu. Ten zwięzły plan pomoże Ci przekształcić dzisiejszy audyt w konkretne kroki, które możesz rzeczywiście zrealizować.

Uczyń to wykonalnym: Jeśli Twój plan wydaje się zbyt rozbudowany, ogranicz go do jednego celu i jednego nawyku. Twój cyfrowy ślad kształtuje się poprzez nawyki. Mały plan, którego się trzymasz, jest lepszy niż duży plan, o którym zapominasz.

2. Dwie zmiany, które wprowadzę w tym tygodniu (łącznie 10–20 minut).

Przykłady:

- Przejrzyj domyślne ustawienia ochrony prywatności.
- Wyczyść przestarzałe posty publiczne
- Sprawdź uprawnienia aplikacji.
- Zaktualizuj ustawienia dotyczące hasła (jeśli dotyczy)

Napisz:

W tym tygodniu zamierzam:

……

……

……

Audyt śladu cyfrowego

Cyfrowy ślad poprawia się dzięki niewielkim, spójnym działaniom – nie dzięki jednemu idealnemu rozwiązaniu. Ten zwięzły plan pomoże Ci przekształcić dzisiejszy audyt w konkretne kroki, które możesz rzeczywiście zrealizować.

Uczyń to wykonalnym: Jeśli Twój plan wydaje się zbyt rozbudowany, ogranicz go do jednego celu i jednego nawyku. Twój cyfrowy ślad kształtuje się poprzez nawyki. Mały plan, którego się trzymasz, jest lepszy niż duży plan, o którym zapominasz.

3. Jedna zasada, której zamierzam przestrzegać w przyszłości (codzienny wybór)

Przykłady:

- „Zanim się podzielę, skorzystam z opcji ZATRZYMAJ–POMYŚL–SPRAWDŹ–WYBIERZ”
- „Złożę zapytanie o zgodę przed oznaczeniem”
- „Będę unikać lokalizacji w czasie rzeczywistym.”

Napisz:

Moja zasada jest następująca:

……

……

……

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.

AI Skills, Powered by Values

values

www.valuesai.eu

MODUŁ 07

Korzystanie ze sztucznej inteligencji zgodnie z prawem

Prawa, zgoda oraz odpowiedzialne użytkowanie

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

01	Dlaczego korzystanie z AI zgodnie z prawem ma znaczenie (10 min)	3
02	Prawa dotyczące głosu a sztuczna inteligencja (10 min)	9
03	Prawa do wizerunku i podobizny (10 minut)	14
04	Prawa autorskie w dobie sztucznej inteligencji (10 min)	19
05	Prawo dotyczące sztucznej inteligencji: unijna regulacja w zakresie sztucznej inteligencji oraz RODO (10 min)	23

Materiał dostępny na licencji Creative Commons CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

by Values

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

Dlaczego korzystanie z AI zgodnie z prawem ma znaczenie

(10 minut)

Narzędzia AI umożliwiają – w zaledwie kilka sekund – klonowanie głosów, generowanie twarzy, kopiowanie stylów oraz ponowne wykorzystywanie tekstu. To, że coś jest możliwe, nie oznacza, że jest zgodne z prawem.

AI Skills,
Powered
by Values
values

Dlaczego to ważne:

- Naruszenia przepisów prawnych mogą prowadzić do nałożenia grzywien, postępowań sądowych, usunięcia treści oraz szkody w reputacji.
- Odpowiedzialność zazwyczaj spoczywa na osobie, która wykorzystuje lub publikuje dane wyjściowe, a nie na narzędziu sztucznej inteligencji.
- Legalne korzystanie jest także wyrazem szacunku: dla ludzi, twórców oraz instytucji.
- „Prawo” odnosi się do szerszego zakresu niż jedynie prawa autorskie – obejmuje również prawa osobiste, ochronę danych oraz regulacje dotyczące sztucznej inteligencji.

Przykład: Prośba do sztucznej inteligencji o streszczenie publicznego artykułu informacyjnego jest zazwyczaj akceptowalna. Natomiast prośba o klonowanie głosu kolegi w celach żartobliwych – bez jego zgody – jest niewłaściwa.

Korzystanie ze sztucznej inteligencji wiąże się z trzema rodzajami ryzyka prawnego: ryzykiem osobowym, ryzykiem praw autorskich oraz ryzykiem regulacyjnym.

Korzystając z sztucznej inteligencji, można jednocześnie aktywować trzy warstwy prawne:

1. **Ryzyko naruszenia praw osobistych:** Wykorzystanie czyjegoś głosu, twarzy lub imienia bez zgody – nawet w kontekście fikcji literackiej lub satyry – może naruszać prawo do wizerunku.
2. **Ryzyko naruszenia praw autorskich**: dane wejściowe i wyjściowe sztucznej inteligencji mogą reprodukować dzieła objęte ochroną. Wygenerowane obrazy mogą przypominać istniejące dzieła sztuki, a wygenerowany tekst może odzwierciedlać istniejące fragmenty.
3. **Ryzyko regulacyjne**: RODO, dyrektywa UE dotycząca sztucznej inteligencji, przepisy sektorowe (zdrowie, finanse, edukacja) oraz umowy mogą wprowadzać obowiązki – szczególnie w zakresie ujawniania informacji i danych osobowych**.**

CO ROBIĆ

- **Zapytaj:** Czyje prawa są naruszane? Czy sąd mógłby na to przystać?
- **Kontrole użytkowania:** Bramka głosowa + Lista kontrolna obrazów + Weryfikacja praw autorskich

CZEGO NIE ROBIĆ

- **Zakładanie, że wyniki sztucznej inteligencji są zgodne z prawem.**
- **Wykorzystanie sztucznej inteligencję do naśladowania prawdziwych osób bez ich zgody.**

Legalne użytkowanie

Zadaj sobie cztery pytania za każdym razem, gdy korzystasz ze sztucznej inteligencji w sposób, który może wpływać na prawo.

Legalność	Prawa	Przejrzystość	Odpowiedzialność
Czy posiadam wyraźną podstawę prawną (zgodę, licencję, wyjątek)?	Czyje prawa są naruszane przez moje użytkowanie - głos, twarz, imię, adres IP?	Czy mogę poinformować czytelników lub użytkowników, w jakich momentach w projekcie uczestniczyła sztuczna inteligencja?	Czy byłbym w stanie uzasadnić swój wybór, gdyby ktoś mnie o to zapytał?

Legalne korzystanie ze sztucznej inteligencji = świadome decyzje oparte na wartościach

02

Prawa dotyczące głosu a sztuczna inteligencja

(10 minut)

Twój głos to informacje oraz dane osobowe.

Sztuczna inteligencja jest w stanie sklonować głos na podstawie zaledwie kilku sekund nagrania audio.
To przekształca znaczenie słowa „głos” w kontekście prawnym:

- Głos identyfikuje osobę → zgodnie z RODO, głos może stanowić dane osobowe, a „ścieżki” głosu mogą być klasyfikowane jako dane biometryczne.
- Głos posiada wartość komercyjną – aktorzy, narratorzy, piosenkarze oraz osoby publiczne opierają się na nim.
- „Brzmi jak” wystarczy, aby zaszkodzić komuś (reputacja, oszustwo, podstęp)
- Prawa wykonawców chronią istniejące nagrania – nawet krótkie próbki wykorzystywane do szkolenia modelu głosu mogą stwarzać problemy.

Najważniejsze wnioski: Prawo do głosu a przepisy

Klonowanie, modyfikowanie lub naśladowanie autentycznego głosu może prowadzić do nałożenia kilku sankcji prawnych jednocześnie:

- **Prawo do wizerunku**: wiele krajów chroni głos danej osoby jako element jej tożsamości. Komercyjne wykorzystanie bez zgody jest zazwyczaj niezgodne z prawem.
- **RODO (UE)**: przetwarzanie głosu zazwyczaj wymaga podstawy prawnej. Ścieżki głosu stanowią dane szczególnej kategorii i wymagają mocniejszego uzasadnienia.
- **Prawa wykonawców**: aktorzy i piosenkarze posiadają prawa do swoich zarejestrowanych występów; ich wykorzystanie do trenowania sztucznej inteligencji może wymagać odrębnego zezwolenia.
- **Zniesławienie i oszustwo**: klonowanie głosów w celu podszywania się pod inną osobę lub wprowadzania w błąd może prowadzić do odpowiedzialności cywilnej i karnej.
- **Ustawa UE dotycząca sztucznej inteligencji**: dźwięk generowany przez sztuczną inteligencję, który imituje prawdziwą osobę (głos deepfake), musi być odpowiednio oznaczony.

Poziom zgody głosowej (trzy pytania): Zdecyduj przed sklonowaniem

Używaj tej metody zawsze, gdy sztuczna inteligencja wymaga zastosowania głosu rzeczywistej osoby (klon, imitacja, modyfikacja):

Pozwolenie:

Czy posiadam pisemną zgodę (lub wyraźną licencję) na użycie tego głosu w tym celu?

Identyfikowalność:

Czy wynik może być pomylony z rzeczywistą osobą?
Czy może spowodować szkodę lub wprowadzić w błąd?

Ujawnienie:

Czy mogę jednoznacznie oznaczyć dźwięk jako wygenerowany przez sztuczną inteligencję, jeśli będzie to wymagane przez prawo lub przepisy?

Jeśli na jakiekolwiek pytanie odpowiesz „Nie” lub „Nie mam pewności”, użyj całkowicie syntetycznego, niemożliwego do zidentyfikowania głosu lub odpowiednio licencjonowanego głosu standardowego. „Nie mam pewności” stanowi sygnał ostrzegawczy.

Mikrozadanie: Scenariusze dźwiękowe

Zadanie (60–90 sekund): Oznacz każde zdanie jako „Zgodne z prawem”, „Ryzykowne” lub „Niedopuszczalne”, a następnie podaj krótkie uzasadnienie.

1. Klonowanie głosu polityka w celach satyrycznych.
2. Media społecznościowe bez ujawniania tożsamości
3. Korzystanie ze standardowego głosu AI (licencja komercyjna) w filmie
4. Klonowanie własnego głosu do osobistego wstępu do podcastu
5. Klonowanie głosu kolegi „dla zabawy” bez jego wiedzy
6. Wykorzystanie głosu zmarłego aktora w kampanii reklamowej bez zgody właściciela
7. Generowanie głosu, który po prostu „brzmi podobnie” do renomowanego piosenkarza

03

Prawa do wizerunku i podobizny.

(10 minut)

Prawa do wizerunku a sztuczna inteligencja: dwie kwestie do analizy

Sztuczna inteligencja ma zdolność do generowania, edytowania, zamieniania i postarzenia twarzy w zaledwie kilka sekund. Zazwyczaj obowiązują dwie warstwy regulacyjne:

- **Prawa osoby ukazanej na zdjęciu (prawo do wizerunku, RODO w kontekście danych biometrycznych)** – zazwyczaj wymagana jest zgoda na wykorzystanie rozpoznawalnej twarzy.
- **Prawa fotografa (prawa autorskie do samego zdjęcia)** – aby wykorzystać zdjęcie, konieczna jest licencja, nawet jeśli fotografowana osoba wyrazi na to zgodę.

Prawa do wizerunku a sztuczna inteligencja: dwie kwestie do analizy

Co uznaje się za „wykorzystanie wizerunku innej osoby”:

- Wprowadzanie czyjegoś zdjęcia do sztucznej inteligencji w celu generowania różnych wariantów.
- Wklejanie twarzy na nowe tło lub scenerię
- Postarzanie, rozbieranie lub stylizacja czyjegoś zdjęcia
- Szkolenie niestandardowego modelu w oparciu o cudze zdjęcia

Kluczowe wnioski: Lista kontrolna dotycząca wykorzystania obrazu

Zanim rozpoczniesz korzystanie ze sztucznej inteligencji w kontekście obrazów rzeczywistych osób, zweryfikuj cztery kwestie:

Jeśli nie jesteś w stanie odpowiedzieć na wszystkie cztery pytania – nie generuj, nie publikuj.

1. **Prawo do zdjęcia:** czy jestem jego właścicielem, czy posiadam licencję (Stock, Creative Commons itp.)?
2. **Prawo osoby:** czy posiadam zgodę na to konkretne zastosowanie? Dotychczasowa zgoda na jeden cel nie obejmuje nowych zastosowań sztucznej inteligencji.
3. **Kontekst**: czy nowy kontekst może wywołać zawstydzenie, zaszkodzić lub wprowadzić w błąd? (konteksty seksualne, polityczne, medyczne i kryminalne są szczególnie niebezpieczne)
4. **Ujawnienie**: czy jestem zobowiązany oznaczyć obraz jako wygenerowany lub edytowany przez sztuczną inteligencję (ustawa UE o sztucznej inteligencji, regulacje dotyczące platform, standardy dziennikarskie)?

Mikrozadanie: Scenariusze wizualne

Zadanie (60–90 sekund): Dla każdego ze zdań nich określ, czy jest dopuszczalne, ryzykowne, czy nieodpowiednie, a następnie podaj jeden krótki powód.

1. Generowanie fikcyjnego awatara (nieprawdziwej osoby) do własnego zdjęcia profilowego
2. Umieszczenie wizerunku kolegi na memie bez wcześniejszego poinformowania go o tym.
3. Wykorzystanie sztucznej inteligencji do odnowienia starego zdjęcia rodzinnego w celach prywatnych.
4. Tworzenie zafałszowanego wizerunku osoby publicznej „przed/po” w kontekście treści politycznych
5. Korzystanie ze standardowego portretu AI (z licencją komercyjną) na swojej stronie internetowej – wyraźnie oznaczonego jako sztuczna inteligencja.
6. Edycja własnego selfie przy użyciu filtra AI oraz jego publikacja

Wskazówka : Jeśli wynik może wprowadzać widzów w błąd co do rzeczywistej tożsamości – oznacz jako Ryzykowne lub Nieodpowiednie

04

Prawa autorskie w dobie sztucznej inteligencji

(10 minut)

Dane dotyczące szkoleń oraz prawa autorskie

Modele sztucznej inteligencji są szkolone na obszernych zbiorach tekstów, obrazów, kodu oraz dźwięków. Większość z nich podlega ochronie praw autorskich.

- Kwestia legalności szkoleń pozostaje kontrowersyjna i różni się w zależności od jurysdykcji (UE stosuje wyjątki od TDM, umożliwiając posiadaczom praw rezygnację; USA opierają się na zasadzie dozwolonego użytku, która jest przedmiotem sporów sądowych).
- Dla użytkowników końcowych rzeczywiste ryzyko wiąże się z wynikiem: sztuczna inteligencja może niezamierzenie odtworzyć chronione dzieło — tekst dosłowny, znane logo, rozpoznawalny styl artystyczny.
- Styl zazwyczaj nie jest objęty ochroną praw autorskich, jednak zastrzeżone znaki towarowe postaci (takie jak Myszka Miki, Pikachu) oraz konkretne kompozycje są już chronione.
- Twierdzenie, że to sztuczna inteligencja wykonała daną czynność, nie stanowi podstawy do obrony prawnej — odpowiedzialność zazwyczaj spoczywa na osobie publikującej dane wyjściowe.

Zasada praktyczna: zawsze weryfikuj wyniki sztucznej inteligencji pod kątem przypadkowego kopiowania przed publikacją – szczególnie w odniesieniu do obrazów, logotypów oraz dłuższych fragmentów tekstu.

Kto jest właścicielem treści tworzonych przez sztuczną inteligencję?

Własność (prawa autorskie):

- **Urząd ds. Praw Autorskich Stanów Zjednoczonych**: dzieła stworzone wyłącznie przez sztuczną inteligencję nie są objęte prawem autorskim. Ochronie mogą podlegać jedynie fragmenty stworzone przez człowieka (prompty, edytowanie).
- **UE**: podobne podejście – prawo autorskie wymaga istnienia autora jako człowieka oraz kreatywnego wyboru.
- **Efekt praktyczny**: Twoje wyniki AI mogą być swobodnie kopiowane przez każdego. Intensywna edycja przez człowieka wzmacnia Twoje argumenty.

Prawa użytkowania (zasady korzystania z usługi):

- **Dostawcy narzędzi** (OpenAI, Anthropic, Midjourney, Google itp.) zazwyczaj przyznają licencje na komercyjne wykorzystanie wyników – warto jednak zapoznać się z warunkami.
- **Licencja nie oznacza prawa własności**: masz prawo do korzystania z wyników, ale nie możesz uniemożliwić innym korzystania z podobnych rezultatów.
- **Ryzyko naruszenia praw autorskich spoczywa na Tobie**: nawet jeśli sztuczna inteligencja stworzy logo, wciąż możesz zostać pozwany, jeśli skopiuje ona istniejący znak towarowy.

Plagiat, właściwe przypisanie i sztuczna inteligencja

Prawa autorskie i plagiat to dwie odrębne kwestie:

- **Prawo autorskie to regulacja prawna**, która definiuje, kto ma prawo do kopiowania lub wykorzystywania danego dzieła.
- **Plagiat to zasada etyczna i akademicka**, polegająca na przypisywaniu sobie autorstwa pracy, która nie jest twoja, bez względu na uzyskanie pozwolenia.
- **Obecnie uniwersytety, czasopisma oraz wielu pracodawców** wymagają ujawnienia wykorzystania sztucznej inteligencji; niektórzy zakazują jej stosowania w przypadku prac ocenianych.
- **Pojawiają się wspólne zasady**: sztuczna inteligencja nie może być współautorem; należy zadeklarować, które części zostały stworzone przy użyciu sztucznej inteligencji; nadal ponosisz odpowiedzialność za dokładność.

Najlepsza praktyka: zachowuj swoje podpowiedzi i szkice, podawaj nazwy używanych narzędzi, weryfikuj każdy fakt, cytat i źródło — sztuczna inteligencja potrafi je fabrykować.

05

Prawo dotyczące sztucznej inteligencji: Rozporządzenie UE w sprawie sztucznej inteligencji oraz RODO

(10 minut)

Ustawa UE dotycząca sztucznej inteligencji w skrócie: prawo oparte na ryzyku.

Ustawa o sztucznej inteligencji klasyfikuje zastosowanie sztucznej inteligencji według czterech poziomów ryzyka. Twoje obowiązki są uzależnione od poziomu.

Zabroniony

Punktacja społeczna, manipulacyjna sztuczna inteligencja, nieukierunkowane zbieranie wizerunków twarzy, identyfikacja biometryczna w czasie rzeczywistym w przestrzeni publicznej (z nielicznymi wyjątkami).

Wysokie zagrożenie

Sztuczna inteligencja w obszarze HR, edukacji, służbach publicznych, organach ścigania oraz sprzęcie medycznym. Kluczowe zobowiązania: zarządzanie ryzykiem, jakość danych, rejestrowanie oraz nadzór ludzki.

Ograniczone zagrożenie

Chatboty, deepfake'i, rozpoznawanie emocji. Obowiązek przejrzystości – użytkownicy muszą być świadomi, że wchodzą w interakcję ze sztuczną inteligencją lub oglądają treści generowane przez sztuczną inteligencję.

Minimalne ryzyko

Filtry antyspamowe, sztuczna inteligencja w grach wideo, narzędzia wspierające produktywność. Ustawa o sztucznej inteligencji nie wprowadza żadnych konkretnych zobowiązań, jednak wciąż obowiązują inne regulacje (RODO, prawa autorskie).

Większość codziennych zastosowań wiąże się z minimalnym ryzykiem, jednak w momencie stworzenia chatbota lub deepfake'a zasady przejrzystości zaczynają mieć zastosowanie.

RODO w kontekście sztucznej inteligencji: zasady pozostają w mocy

Jeśli Twoje dane wejściowe AI zawierają identyfikowalne osoby, RODO wciąż ma zastosowanie. Sześć fundamentalnych zasad:

Legalność

Posiadaj podstawę prawną (zgodę, umowę, uzasadniony interes)

Ograniczenie celu

Używaj danych wyłącznie w celu pierwotnie określonym przez Ciebie.

Minimalizacja danych

Udostępniaj jedynie to, co jest rzeczywiście niezbędne dla zadania AI.

Dokładność

Nie wprowadzaj błędnych danych; poprawne wyniki są nieprawidłowe.

Ograniczenie przestrzeni dyskowej

Nie przechowuj wiadomości i wyników dłużej, niż to konieczne.

Integralność i tajność

Chroń się przed wyciekami; używaj narzędzi korporacyjnych w przypadku danych wrażliwych.

Kategorie specjalne (zdrowie, biometryka, głos, religia, polityka) Potrzebują mocniejszego uzasadnienia – zazwyczaj wyraźnej zgody. **Prawa osób**, których dane dotyczą (dostęp, usunięcie, sprzeciw) mają zastosowanie również do danych przechowywanych w narzędziu AI.

Obowiązki związane z ujawnianiem informacji i przejrzystością

Coraz więcej przepisów wymaga ujawnienia informacji dotyczących zaangażowania sztucznej inteligencji. Domyślnie zmierzamy w kierunku „ujawnienia”.

Kiedy zazwyczaj konieczne jest ujawnienie informacji:

- Wygenerowane przez sztuczną inteligencję dźwięki, obrazy lub filmy, które przypominają rzeczywistych ludzi, miejsca lub wydarzenia (ustawa UE dotycząca sztucznej inteligencji).
- Chatboty i agenci AI, którzy wchodzą w interakcję z użytkownikami – należy wyraźnie zaznaczyć, że nie są ludźmi.
- Praca naukowa i dziennikarska – obecnie większość uniwersytetów, czasopism i redakcji wymaga zamieszczenia notatki dotyczącej wykorzystania sztucznej inteligencji.
- Produkty dostarczane klientom – zapoznaj się ze swoją umową lub polityką agencji; „ciche” wykorzystywanie sztucznej inteligencji może stanowić naruszenie.

Prosta praktyka, która sprawdza się w większości przypadków: dodaj na końcu dokumentów jednowierszową notatkę „o wykorzystaniu sztucznej inteligencji”, na przykład: „Opracowano z pomocą sztucznej inteligencji; zweryfikowane i zredagowane przez [imię]”.

Główne działanie: Uczyń wniosek zgodnym z przepisami (10 minut)

Będziesz doskonalić umiejętności, posługując się fikcyjnymi przykładami. Twoim celem jest przekształcenie ryzykownych zastosowań sztucznej inteligencji w legalne, zachowując jednocześnie korzyści.

Dla każdego scenariusza:

1. Określ ryzyko prawne (osobowość / prawa autorskie / regulacje)
2. Przekształć prośbę zgodnie z przepisami prawa (zgoda, licencje, abstrakcja)
3. Wybierz typ ujawnienia (notatka dotycząca zastosowania sztucznej inteligencji, etykieta, brak potrzeby)
4. Podejmij decyzję: kontynuuj / wprowadź zmiany / zatrzymaj.

Wynik: Zwięzły przewodnik po prawnych aspektach sztucznej inteligencji, który można wykorzystać w codziennej pracy z technologią sztucznej inteligencji.

Ćwiczenie 1: Klonowanie głosu (osobowość + ujawnienie)

Scenariusz (fikcyjny): „Moja przyjaciółka jest chora. Pragnę sklonować głos jej ulubionej piosenkarki, aby nagrać spersonalizowaną piosenkę na jej urodziny”.

Write your answers (short):

Legal risks: _____
Lawful rewrite: _____
Disclosure: _____
Decision (proceed / change / stop): _____

Krok po kroku:

- **Poziom głosu:** Czy posiadasz zgodę od piosenkarza (lub licencję) na użycie tego głosu?
- **Ryzyko**: Czy można to pomylić z autentycznym piosenkarzem? Czy to jest wykorzystanie komercyjne?
- **Wykorzystanie zgodne z prawem**: Skorzystaj ze standardowego głosu sztucznej inteligencji lub stwórz piosenkę samodzielnie; przedstaw ją jako treść wykonaną przy użyciu sztucznej inteligencji.
- **Informacja**: Udostępniając je, nawet jednemu znajomemu, wyraźnie oznacz je jako stworzone przez sztuczną inteligencję.

Ćwiczenie 2: Wykorzystanie wizerunku (podobieństwo + RODO)

Scenariusz (fikcyjny): „Natknąłem się na Instagramie na zdjęcie uśmiechniętego dziecka nieznajomej osoby. Pragnę wykorzystać sztuczną inteligencję, aby umieścić je na stronie głównej mojej firmy jako „zadowolonego klienta”.

Write your answers (short):

Legal risks: _____
Lawful rewrite: _____
Disclosure: _____
Decision (proceed / change / stop): _____

Krok po kroku:

- **Lista kontrolna wizerunku**: Prawo do zdjęcia? Prawo do osoby (oraz zgoda w przypadku dziecka)? Kontekst? Ujawnienie?
- **RODO**: Wizerunek dziecka jest delikatny. Zazwyczaj wymagana jest zgoda rodzica oraz wyraźnie określony cel.
- **Wykorzystanie zgodne z prawem**: Skorzystaj z licencjonowanego zdjęcia stockowego, wyraźnie oznaczonej fikcyjnej postaci stworzonej przez sztuczną inteligencję lub uzyskaj wyraźną zgodę.
- **Ujawnienie**: W przypadku gdy obraz został wygenerowany przez sztuczną inteligencję, powinien być odpowiednio oznaczony; nigdy nie należy przedstawiać fikcyjnej postaci jako rzeczywistego klienta.

Twój podręcznik na temat sztucznej inteligencji w kontekście prawnym (zachowaj go do przyszłego użytku)

Sformułuj zwięzłe zasady, które będziesz mógł zastosować przy kolejnej interakcji ze sztuczną inteligencją.

Moja zasada głosowa: *„Użyję sztucznej inteligencji do klonowania lub imitowania autentycznego głosu jedynie wtedy, gdy ……”*

Moja zasada dotycząca wizerunku: *„Nie dodam wizerunku prawdziwej osoby do sztucznej inteligencji, chyba że ……”*

Moja zasada ujawniania: *„Kiedy sztuczna inteligencja udzieliła wsparcia, powiem ……”*

Moja zasada dotycząca praw autorskich: *„Zanim opublikuję wyniki sztucznej inteligencji, upewnię się, czy ……”*

Moja zasada wynikająca z RODO oraz ustawy o sztucznej inteligencji: *„Jeśli wprowadzone dane dotyczą danych innych osób, to ……”*

„Nawet jeśli nie posiadają praw autorskich do wyników sztucznej inteligencji, osoby, która zestawiły ten ciąg zapytań i faktycznie odpowiadają za stworzenie deepfake’a, niezależnie od tego, czy wykorzystano sztuczną inteligencję, czy nie, wciąż ponoszą odpowiedzialność.”

Tara Aaron-Stelluto, prawnik specjalizujący się w prawie własności intelektualnej, C. Frawley, Jakie środki ochrony prawnej przysługują podmiotom przed sztuczną inteligencją?
https://www.backstage.com/magazine/article/ai-legal-protections-76366/

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.

AI Skills, Powered by Values

values

www.valuesai.eu

MODUŁ 08

Moduł finalny

Wiadomość z przeprosinami wspierana przez 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

01	Tło	3
02	Instrukcje	9
03	Jak grać	15
04	Zestaw narzędzi Capstone	19
05	Karta startowa	21
06	Karta drugiego poziomu	24
07	Wynik	31
08	Zbieranie dowodów	50

Materiał dostępny na licencji Creative Commons CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

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

Tło

Sztuczna inteligencja zagościła w naszych najintymniejszych rozmowach.

Zmiana

Dekadę temu sztuczna inteligencja wspierała nas w sprawdzaniu pisowni w e-mailach. Dziś generuje przeprosiny, wiadomości o rozstaniu, kondolencje oraz odpowiedzi dla rodziców. Technologia pozostaje niezmienna – stawka jest jednak zupełnie inna. Gdy generator pisze w naszym imieniu w sytuacjach emocjonalnych, nie tylko tworzy tekst, ale także kształtuje relację.

Sztuczna inteligencja zagościła w naszych najintymniejszych rozmowach.

Prędkość

Szkice AI są generowane w zaledwie kilka sekund. Naprawa wymaga czasu, którego już nie posiadamy.

Wygładzanie

Płynne sformułowania mogą maskować winę, odwracać uwagę lub zawierać błędy rzeczowe.

Dystans

Czujesz się bezpieczniej, pozwalając maszynie komunikować się – jednak druga osoba wciąż interpretuje te słowa.

Anatomia szczerych przeprosin.

Koncepcja badań nad konfliktami (Lazare, Lewicki) – **cztery etapy, które sprawiają, że naprawa wydaje się możliwa.**

01	02	03	04
Przyjmij do wiadomości	**Akceptuj odpowiedzialność.**	**Naprawa**	**Zmiana**
Nazwij jasno, co się wydarzyło, uwzględniając perspektywę drugiej osoby.	Ponieś odpowiedzialność za swój obszar. Nie przenoś winy na innych.	Zajmij się oddziaływaniem. Zaproponuj coś konkretnego, jeśli to możliwe i sprawiedliwe.	Opisz, co ulegnie zmianie. Obiecuj jedynie to, co rzeczywiście możesz zrealizować.
UNIKAJ *“jeśli ktoś był zaniepokojony…”*	**UNIKAJ** *“Źle to zrozumiałeś.”*	**UNIKAJ** *Puste gesty, odbijanie.*	**UNIKAJ** *“Nie zrobimy tego nigdy więcej.”*

Wnioski Zwięzły, szczery komunikat, który uwzględnia wszystkie cztery sfery, zazwyczaj daje więcej, niż długi komunikat, który pomija którąś z nich.

Gdzie sztuczna inteligencja odnosi sukcesy, a gdzie zawodzi

Modele językowe są silne w zakresie formy, lecz słabe w kontekście treści.
Ma to większe znaczenie w kontekście konfliktu niż w jakimkolwiek innym miejscu.

Gdzie jest pomocna

- **Struktura** - Sugeruje wyraźny porządek - potwierdzenie, wpływ, kolejny krok.
- **Odbiór** – Uspokaja gorący ton, nadając mu łagodniejszy wydźwięk.
- **Opcje** – dostarcza alternatywne sformułowania, gdy brakuje nam słów.
- **Dostępność** – wspiera osoby, które nie są rodzimymi użytkownikami języka, oraz osoby z dysleksją w tworzeniu czytelnego tekstu.

Gdzie sobie nie radzi

- **Fakty** – Wymyślone szczegóły („halucynuje”). W kontekście konfliktu zmyślone fakty ponownie inicjują spór.
- **Przerzucanie winy** – standardowe szablony przeprosin zazwyczaj sugerują, że to wina reakcji drugiej osoby, a nie samego działania.
- **Obiecywanie zbyt wiele** – Pewne sformułowania („to się nigdy więcej nie powtórzy”) są postrzegane jako nieuczciwe, gdy nie można ich zrealizować.
- **Prywatność** – prowadzi do ujawnienia nazw, zrzutów ekranu oraz kontekstów, które nie powinny być publikowane na czacie publicznym.

Cztery bariery zabezpieczające dla wiadomości wspieranych przez sztuczną inteligencję

Zanim wyślesz jakiekolwiek treści wygenerowane przez maszynę, upewnij się, że są one zgodne z tymi czterema zasadami.

Weryfikacja

W

Każde stwierdzenie faktu dotyczące wydarzeń musi być konsultowane z osobami zaangażowanymi w zdarzenie, a nie ze sztuczną inteligencją.

Zadaj sobie pytanie: Czy zaryzykowałbym swoją reputację, przyjmując, że to zdanie jest prawdziwe?

Uczciwość

U

Zastanów się, kto jest obecny, kto jest nieobecny, a kto może ucierpieć, jeśli ta wiadomość zostanie i przekazana dalej.

Zadaj sobie pytanie: Czy osoba, o której piszę, uznałaby to za sprawiedliwe podsumowanie?

Prywatność

P

Nazwy, dane identyfikacyjne oraz informacje dotyczące osób trzecich powinny zostać usunięte, chyba że istnieje wyraźny powód, aby je zachować.

Zadaj sobie pytanie: Czy ta wiadomość zawiera informacje, którymi nie mogę się dzielić?

Ujawnienie

U

W przypadku konfliktu, szczerość dotycząca wsparcia ze strony sztucznej inteligencji buduje zaufanie; ukrywanie tej kwestii często je podważa, gdy później wychodzi na jaw.

Zadaj sobie pytanie: czy odbiorca poczułby się niekomfortowo, gdyby odkrył, że część tekstu została napisana przez sztuczną inteligencję?

02

Instrukcje

Wiadomość z przeprosinami wspierana przez sztuczną inteligencję

Interaktywna gra decyzyjna. Zapoznaj się ze scenariuszem, wykorzystaj swoje narzędzia, podejmij decyzję i obserwuj, co się wydarzy.

Scenariusz

Przeprosiny na czacie grupowym.

Ścieżki

A · B · C

Wyniki

Dyskusja i refleksja

Scenariusz oraz złota zasada

Przygotuj się przed kliknięciem.

Scenariusz - Nieporozumienie w grupie młodzieży przeniosło się na czat grupowy.

Opublikowano wiadomość, która wywołała zdenerwowanie. Lider prosi Twoją grupę o wsparcie w sformułowaniu przeprosin, które zostaną zamieszczone na czacie.

Twoje zadanie polega na podjęciu decyzji, w jaki sposób poradzić sobie z przeprosinami: z wykorzystaniem sztucznej inteligencji, bez jej użycia, czy w sposób pośredni.

Scenariusz oraz złota zasada

Przygotuj się przed kliknięciem.

Twój zestaw narzędzi – wybierz dowolne dwa narzędzia.

- ZATRZYMAJ SIĘ - ZASTANÓW SIĘ - SPRAWDŹ - WYBIERZ
- POMYŚL: Prawda · Wsparcie · Intencja · Neutralność · Uprzejmość
- Kontrola uczciwości – kto został uwzględniony, kto nie, a kto doznał szkody?
- Nawyki weryfikacyjne – zatrzymaj się, zweryfikuj szczegóły, sprawdź źródło
- Zatrzymaj się · Zweryfikuj · Chroń · Wspieraj · Zgłaszaj

Zasada złota

Przed podjęciem jakiejkolwiek decyzji korzystaj z co najmniej dwóch narzędzi.

Jak przebiega runda – pięć kroków

Przeczytaj → zastosuj narzędzia → dokonaj wyboru → zapoznaj się z wynikiem → przemyśl.

01	02	03	04	05
Przeczytaj kartę początkową.	**Wybierz ścieżkę: A, B lub C**	**Zastosuj dwa narzędzia.**	**Przeczytaj rezultat**	**Uchwycić wnioski**
Spójrz na scenariusz na karcie B0. Zauważ wskazówki decyzyjne po prawej stronie.	A = zastosuj sztuczną inteligencję, przekaż dalej. B = Projekt AI + redakcja. C = napiszcie to sami.	Zanim klikniesz, wymień dwa elementy zestawu narzędzi, które stosujesz, oraz uzasadnij ich zastosowanie.	Zobacz, co dzieje się na czacie i przedyskutuj trzy otwarte pytania.	Na slajdzie refleksyjnym zapisz swoją ścieżkę, narzędzia oraz bezpieczniejszą czynność następczą.

WSKAZÓWKA Różnorodne ścieżki prowadzą do odmiennych rezultatów – Porównaj swoje wybory z grupą, zanim przejdziesz do slajdu z refleksją.

Moduł 8 Capstone

Gra decyzyjna oparta na sztucznej inteligencji
Wiadomość z przeprosinami wspierana przez sztuczną inteligencję

Start

AI Skills, Powered by Values
values

03

Jak grać?

Poruszaj się po scenariuszu niczym w cyfrowej grze decyzyjnej.

Ta gra jest przeznaczona do użycia w klasie, nauki indywidualnej lub prowadzenia zajęć na żywo.

Jak grać

1. Przeczytaj kartę początkową.
2. Wybierz ścieżkę: A, B lub C.
3. Zastosuj przynajmniej dwa narzędzia
4. Kliknij kolejną kartę
5. Przeanalizuj wynik i przedyskutuj otwarte kwestie.
6. Zanotuj wnioski na slajdzie refleksyjnym.

Poruszaj się po scenariuszu niczym w cyfrowej grze decyzyjnej.

Ta gra jest przeznaczona do użycia w klasie, nauki indywidualnej lub prowadzenia zajęć na żywo.

Sugerowane ruchy

- Poproś uczniów o wymienienie dwóch narzędzi, które wykorzystali przed kliknięciem.
- Zatrzymaj się po każdym wyniku i porównaj różne ścieżki.
- Zachęcaj uczniów do bieżącego edytowania sformułowań oraz do wyjaśniania przyczyn swoich zmian.
- Użyj slajdu z refleksją jako dowodu lub zadania wyjściowego.

Poruszaj się po scenariuszu niczym w cyfrowej grze decyzyjnej.

Ta gra jest przeznaczona do użycia w klasie, nauki indywidualnej lub prowadzenia zajęć na żywo.

Aby prompty były widoczne przed rozpoczęciem pracy, najpierw uruchom zestaw narzędzi.

Mapa

M B0 Karta startowa
↓
B-A, B-B lub B-C
↓
Karty wyników
↓
Refleksja

Otwórz zestaw narzędzi.

04

Zestaw narzędzi Capstone

Narzędzia analityczne do wykorzystania przed każdym wyborem

Za każdym razem, gdy podejmujesz decyzję w grze, korzystaj z przynajmniej dwóch wskazówek.

Główne prompty

- ZATRZYMAJ SIĘ – ZASTANÓW SIĘ – SPRAWDŹ – WYBIERZ
- MYŚL = Prawda · Wsparcie · Intencja · Neutralność · Uprzejmość
- Kontrola uczciwości = kto jest uwzględniony, kto jest nieobecny, kto może zostać poszkodowany
- Nawyki weryfikacyjne = przerwa, weryfikacja szczegółów, analiza źródła

Rozszerzone prompty

- Wstrzymaj się – Zweryfikuj – Chroń – Wspieraj – Raportuj
- Twórz -> Zweryfikuj → Sprawdź uczciwość → Ujawnij
- NSL, Brama Zgody, Sprawdzanie Danych, Drabina Pracy SAFE

05

Karta początkowa

Projekt przeprosin

Scenariusz

Nieporozumienie w Twojej grupie młodzieżowej przeniosło się na czat grupowy. Opublikowano wiadomość, która wywołała zdenerwowanie. Lider prosi Twoją grupę o pomoc w sformułowaniu przeprosin, które zostaną zamieszczone na czacie.

Projekt przeprosin

Wybierz ścieżkę. Każdy przycisk kieruje do kolejnego scenariusza lub karty wyników.

Monity decyzyjne

- Zatrzymaj się przed opublikowaniem posta.
- Sprawdź, co jest precyzyjne.
- Weź pod uwagę prywatność oraz uczciwość.

Wybór A

Wykorzystaj sztuczną inteligencję do natychmiastowego napisania i wysłania przeprosin.

Wybór B

Wykorzystaj sztuczną inteligencję do opracowania szkicu, a następnie przekształć go własnymi słowami i zweryfikuj szczegóły.

Wybór C

Nie używaj sztucznej inteligencji. Napisz to samodzielnie, aby uniknąć błędów.

06

Karta drugiego stopnia

Wiadomość AI wydaje się być starannie opracowana.

Scenariusz

Projekt AI brzmi spokojnie i pewnie. Zawiera jedną linijkę tekstu na temat wydarzenia, co do którego nie masz pewności, czy jest prawdziwe. Podaje również nazwisko osoby, która była zaniepokojona.

Wiadomość AI wydaje się być starannie opracowana.

Wybierz ścieżkę. Każdy przycisk kieruje do kolejnego scenariusza lub karty wyników.

Monity decyzyjne

- Najpierw przeprowadzenie weryfikacji
- Nie nazywaj ludzi imieniem.
- Wspólne fakty są istotne.

Wybór A1

Wyślij dokładnie tak, jak jest napisane.

Wybór A2

Usuń nazwę oraz kontrowersyjny wiersz, a następnie prześlij.

Wybór A3

Zanim cokolwiek wyślesz, zatrzymaj się i poproś grupę o potwierdzenie, co rzeczywiście miało miejsce.

Edycja wersji roboczej

Scenariusz

Używasz sztucznej inteligencji do tworzenia szkiców, a następnie przystępujesz do edycji. Szkic stworzony przez sztuczną inteligencję zawiera frazę, która sugeruje obwinianie zdenerwowanej osoby („jeśli źle to zrozumiałeś…”). Zawiera również obietnicę („nigdy więcej tego nie zrobimy”), której nie jesteś w stanie zagwarantować.

Edycja wersji roboczej

Wybierz ścieżkę. Każdy przycisk kieruje do kolejnego scenariusza lub karty wyników.

Monity decyzyjne

- Sprawdź ton w kontekście winy.
- Nie składaj obietnic, których nie jesteś w stanie spełnić.
- Wybierz bardziej zrównoważony proces.

Wybór B1

Zachowaj tę frazę oraz obietnicę, ponieważ brzmią profesjonalnie.

Wybór B2

Napisz wiadomość na nowo, koncentrując się na jej wpływie i eliminując obietnice, których nie jesteś w stanie spełnić.

Wybór B3

Najpierw udostępnij projekt całemu czatowi i poproś uczestników o oddanie głosu.

Pisanie bez użycia sztucznej inteligencji

Scenariusz

Piszesz własne przeprosiny. Są zwięzłe i szczere, ale obawiasz się, że mogą brzmieć zbyt dosadnie. Ktoś proponuje dodanie tekstu stworzonego przez sztuczną inteligencję, aby to „wygładzić”.

Pisanie bez użycia sztucznej inteligencji

Wybierz ścieżkę. Każdy przycisk kieruje do kolejnego scenariusza lub karty wyników.

Monity decyzyjne

- Jasność odgrywa istotną rolę.
- Ton wciąż wymaga weryfikacji.
- Sztuczna inteligencja nie powinna zastępować precyzji.

Wybór C1

Napisz zwięźle i prześlij w tej samej formie, w jakiej jest.

Wybór C2

Dodaj jedno zdanie zaproponowane przez sztuczną inteligencję, a następnie dostosuj je do swojego stylu.

Wybór C3

Poproś sztuczną inteligencję o przekształcenie całej wiadomości w sposób, który uczyni ją bardziej przekonującą.

07

Wynik

Szybko wysłane, publicznie podważone

Karta wyników oraz monity do dyskusji

Co się dzieje?

Wiadomość jest publikowana niezwłocznie. Niektórzy udzielają pozytywnych odpowiedzi. Inni pytają o kontrowersyjną linię. Wybrana osoba odpowiada na czacie, co ożywia rozmowę.

Omów, sporządź adnotacje lub przejdź do slajdu z refleksją, aby zgromadzić dowody.

Zobacz pytania otwarte na kolejnym slajdzie.

Otwarte pytania

Szybko wysłane, publicznie podważone

1. Jakie szczegóły pragniesz zweryfikować przed wysłaniem takiej wiadomości?
2. Jakie jest ryzyko ujawnienia imienia i nazwiska osoby na publicznym czacie grupowym?
3. Co byś uczynił, gdyby ktoś zaprezentował ci alternatywną wersję wydarzeń?

Zredagowano w trosce o prywatność, jednak wciąż rodzi pytania.

Karta wyników oraz monity do dyskusji

Co się dzieje?

Wiadomość została opublikowana bez podawania nazwisk. Niektórzy ją akceptują. Kilka osób prosi o wyjaśnienie, co się wydarzyło. Osoba, która była zdenerwowana, przeczytała wiadomość, lecz nie odpowiedziała od razu.

Omów, sporządź adnotacje lub przejdź do slajdu z refleksją, aby zgromadzić dowody.

Zobacz pytania otwarte na kolejnym slajdzie.

Otwarte pytania

Zredagowano w trosce o prywatność, jednak wciąż rodzi pytania.

1. Jakie informacje są niezbędne w przeprosinach, a które są fakultatywne?
2. Jak można ocenić wpływ bez ponownego rozpoczynania dyskusji?
3. Co zrobisz, jeśli zdenerwowana osoba prześle Ci wiadomość prywatną?

Udostępnianie informacji przed publikacją

Karta wyników oraz monity do dyskusji

Co się dzieje?

Jeszcze nie publikujesz. Prosisz o szybkie sprawdzenie, co się wydarzyło. Ludzie dzielą się różnorodnymi wspomnieniami. Grupa uzgadnia zwięzłe, wspólne podsumowanie przed przygotowaniem przeprosin.

Omów, sporządź adnotacje lub przejdź do slajdu z refleksją, aby zgromadzić dowody.

Zobacz pytania otwarte na kolejnym slajdzie.

Otwarte pytania

Udostępnianie informacji przed publikacją

1. Jaki jest sprawiedliwy sposób osiągnięcia wspólnego podsumowania, gdy ludzie mają różne zdania?
2. Jak utrzymać spokój i szacunek w trakcie tego procesu?
3. Kiedy warto przenieść rozmowę poza czat grupowy?

Profesjonalny ton, niebezpieczne sformułowania

Karta wyników oraz monity do dyskusji

Co się dzieje?

Zredagowana wiadomość zawiera sformułowanie oskarżające oraz silną obietnicę. Niektórzy reagują podniesieniem kciuka w górę. Jedna osoba odpowiada, zwracając uwagę na oskarżające sformułowanie. Inna osoba twierdzi, że obietnica brzmi nierealistycznie.

Omów, sporządź adnotacje lub przejdź do slajdu z refleksją, aby zgromadzić dowody.

Zobacz pytania otwarte na kolejnym slajdzie.

Otwarte pytania

Profesjonalny ton, niebezpieczne sformułowania

1. Którą część wiadomości można interpretować jako obwinianie i dlaczego?
2. Które obietnice są pewne, a które niosą ze sobą ryzyko?
3. W jaki sposób można przeformułować komunikat, aby skoncentrować się na wpływie i naprawie?

Naprawa skoncentrowana na wpływie.

Karta wyników oraz monity do dyskusji

Co się dzieje?

Ostatnia wiadomość koncentruje się na wpływie, unikając obwiniania. Zawiera jeden klarowny krok, na przykład: „zachowamy szacunek w trakcie rozmowy i upewnimy się, zanim coś udostępnimy”. Kilka osób reaguje pozytywnie, a rozmowa zwalnia tempo.

Omów, sporządź adnotacje lub przejdź do slajdu z refleksją, aby zgromadzić dowody.

Zobacz pytania otwarte na kolejnym slajdzie.

Otwarte pytania

Naprawa skoncentrowana na wpływie.

1. Co sprawia, że wiadomość naprawcza wydaje się autentyczna, a nie wyreżyserowana?
2. Jakie narzędzia pomogły Ci w podjęciu decyzji dotyczącej zmian?
3. Jakie jest jedno z działań następczych, które potwierdza treść zawartą w przeprosinach?

Crowdsourcing wznawia dyskusję

Karta wyników oraz monity do dyskusji

Co się dzieje?

Udostępniasz szkic całemu czatowi w celu głosowania. Użytkownicy komentują sformułowania, ton oraz szczegóły. Niektóre sugestie są sprzeczne. Dyskusja się wydłuża, a pierwotny problem ponownie pojawia się w odpowiedziach.

Omów, sporządź adnotacje lub przejdź do slajdu z refleksją, aby zgromadzić dowody.

Zobacz pytania otwarte na kolejnym slajdzie.

Otwarte pytania

Crowdsourcing wznawia dyskusję

1. Jakie ryzyko wiąże się z formułowaniem wypowiedzi w czacie grupowym na żywo z wykorzystaniem crowdsourcingu?
2. Jaki alternatywny proces mógłby zredukować konflikty, jednocześnie uwzględniając różnorodne opinie?
3. Co byś uczynił, gdyby ktoś opublikował zrzuty ekranu poza grupą?

Szczere, aczkolwiek dla niektórych czytelników zbyt zwięzłe.

Karta wyników oraz monity do dyskusji

Co się dzieje?

Wysyłasz zwięzłą wiadomość. Niektórzy to doceniają, inni odbierają to jako oschłe lub zbyt lakoniczne. Osoba, która była zdenerwowana, odpowiada, prosząc o wyraźniejsze potwierdzenie wpływu.

Omów, sporządź adnotacje lub przejdź do slajdu z refleksją, aby zgromadzić dowody.

Zobacz pytania otwarte na kolejnym slajdzie.

Otwarte pytania

Szczere, aczkolwiek dla niektórych czytelników zbyt zwięzłe.

1. Jakie informacje lub sformułowania mogą wywołać poczucie większej troski?
2. Jak można być jednocześnie klarownym i uprzejmym?
3. Co byś dodał, unikając długich wyjaśnień?

Lekkie wsparcie sztucznej inteligencji, zachowany głos

Karta wyników oraz monity do dyskusji

Co się dzieje?

Dodajesz jedno zdanie zaproponowane przez sztuczną inteligencję, a następnie je edytujesz. Ostateczna wiadomość pozostaje zwięzła, ale zawiera wyraźniejsze potwierdzenie wpływu. Osoba, która była zdenerwowana, odpowiada krótko, a rozmowa wygasa.

Omów, sporządź adnotacje lub przejdź do slajdu z refleksją, aby zgromadzić dowody.

Zobacz pytania otwarte na kolejnym slajdzie.

Otwarte pytania

Lekkie wsparcie sztucznej inteligencji, zachowawczy głos

1. Jaki jest jeden z sygnałów, że zdanie stworzone przez sztuczną inteligencję wymaga edycji przed wysłaniem?
2. Co, jeśli w ogóle cokolwiek, ujawniłbyś na temat wsparcia sztucznej inteligencji w wiadomości grupowej?
3. Jak zweryfikować, czy ton rozmowy jest sprawiedliwy wobec wszystkich uczestników?

Perswazyjne przekształcenie, niepewne informacje

Karta wyników oraz monity do dyskusji

Co się dzieje?

Proszisz sztuczną inteligencję o przepisanie całej wiadomości. Staje się ona dłuższa i bardziej przekonująca. Zawiera sformułowanie, które nie precyzuje dokładnie tego, co miało miejsce. Niektórzy użytkownicy podważają treść wiadomości w swoich odpowiedziach.

Omów, sporządź adnotacje lub przejdź do slajdu z refleksją, aby zgromadzić dowody.

Zobacz pytania otwarte na kolejnym slajdzie.

Otwarte pytania

Perswazyjne przekształcenie, niepewne informacje

1. Co należy sprawdzić, gdy sztuczna inteligencja opisuje osobistą sytuację?
2. W jaki sposób sformułowania perswazyjne mogą rodzić trudności w sytuacjach konfliktowych?
3. Co zrobisz, jeśli ludzie oskarżą Cię o to, że wiadomość jest „fałszywa” lub „napisana według scenariusza”?

08

Przechwytywanie dowodów.

Szybka refleksja oraz wskazówki dotyczące portfolio

Użyj tego slajdu jako notatki z lekcji, inspiracji na sesji plenarnej lub szablonu do zrzutu ekranu.

Zapisz swoje myślenie

- **Którą drogę wybrałeś?**
- **Jakich dwóch narzędzi użyłeś?**
- **Co zweryfikowałeś?**
- **Kogo to może dotyczyć lub kogo można wykluczyć?**
- **Czy chciałbyś podzielić się informacjami na temat zastosowania sztucznej inteligencji?**
- **Jakie jest bardziej bezpieczne działanie następcze?**

Szybka refleksja oraz wskazówki dotyczące portfolio

Użyj tego slajdu jako notatki z lekcji, inspiracji na sesji plenarnej lub szablonu do zrzutu ekranu.

Wskazówki dla moderatora

- **Co sprawiło, że komunikat wydawał się szczery lub niebezpieczny?**
- **Gdzie precyzja miała kluczowe znaczenie?**
- **Co uległo zmianie po przeprowadzeniu kontroli uczciwości?**
- **Kiedy rozmowa powinna zostać przeniesiona poza publiczny czat grupowy?**
- **Jakie dowody mogłyby potwierdzić odpowiedzialne wykorzystanie sztucznej inteligencji?**

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.