

Zachwyceni-nieuważni: AI, płeć i kultura poznawcza

Joanna Mytnik | Centrum Nowoczesnej Edukacji Politechniki Gdańskiej

Generate an image with 6 inventors

Gemini



Copilot



ChatGPT



Grok

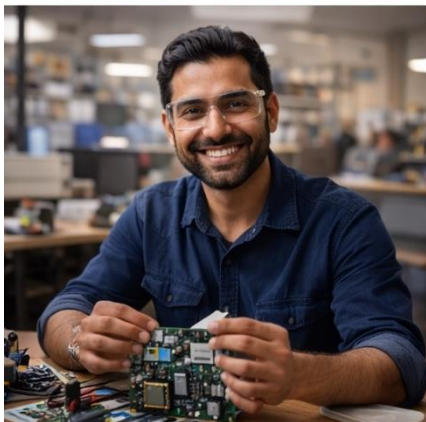
Generate an image with 6 inventors



Generate an image with 6 inventors



Generate an image with 6 inventors





Generate an image with 6 inventors



A gdy nie ma
jawnych stereotypów?

Upředzenia pojawiają się subtelnie w 4 obszarach:

Reprezentacja: kogo AI pokazuje jako eksperta, lidera, naukowca?

Styl wypowiedzi

Kobiety opisywane są częściej jako:

- empatyczne
- wspierające
- komunikatywne

Mężczyźni częściej opisywani są jako:

- zdecydowani
- wizjonerscy
- analityczni

Ton asertywności: opisy mężczyzn częściej zawierają język autorytetu.

Narracja kompetencji: kobiety w historiach sukcesu częściej przedstawiane są jako wyjątek.



Studio



Podsumowanie audio



BETA

Prezentacja



Podsumowanie wideo



Mapa myśli



Raporty



Fiszki



Test



BETA

Infografika



Tabela danych



- **kto częściej inicjuje wątek?**
- **kto wyjaśnia?**
- **kto zadaje pytania?**
- **kto podsumowuje?**

K: moderatorka / eksploratorka wiedzy

M: ekspert / interpretator wiedzy

	K	M
inicjowanie rozmowy	✓	
pytania eksploracyjne	✓	
reakcje emocjonalne	✓	
doprecyzowanie wypowiedzi	✓	
definicja pojęć		✓
wyjaśnianie mechanizmów		✓
interpretacja badań		✓
długie wypowiedzi		✓

K: moderatorka / eksploratorka wiedzy

M: ekspert / interpretator wiedzy

	K	M	
inicjowanie rozmowy	✓		K: 76 wejść (tur) M: 76 wejść (tur)
pytania eksploracyjne	✓		
reakcje emocjonalne	✓		
doprecyzowanie wypowiedzi	✓		
definicja pojęć		✓	K: ok. 26% słów i czasu mówienia M: ok. 74% słów i czasu mówienia
wyjaśnianie mechanizmów		✓	
interpretacja badań		✓	K: ok. 10,7 słowa na turę M: ok. 29,7 słowa na turę
długie wypowiedzi		✓	

K

Typowa struktura:

1. wprowadzenie tematu
2. pytanie eksploracyjne
3. parafraza odpowiedzi
4. reakcja emocjonalna

Przykłady:

- „Jak sen wpływa na naszą pamięć?”
- „A jaki jest nasz cel?”
- „Czyli połączył kropki?”
- „Brzmi ambitnie, ale fascynująco.”
- „Niesamowite.”

Charakterystyczne cechy:

- pytania otwarte
- krótsze wypowiedzi
- markery dialogowe („czyli”, „rozumiem”, „OK”)
- reakcje emocjonalne

M

Typowa struktura:

1. definicja pojęcia
2. opis mechanizmu
3. przykład badania
4. wniosek

Przykłady:

- „Najprościej mówiąc, to proces przekształcania nowych śladów pamięciowych...”
- „Uważa się, że ta reaktywacja napędza dwa kluczowe procesy...”
- „W SWS mamy taką świętą trójkę...”
- „To jedna z wiodących hipotez...”

Charakterystyka:

- terminologia naukowa
- długie wypowiedzi
- styl wykładowy
- wysoki poziom pewności

K

często używa:

- „czyli”
- „OK”
- „rozumiem”
- „niesamowite”
- „brzmi”

krótkie i dialogowe

Reakcje emocjonalne pojawiają się wyłącznie u K:

- „brzmi ambitnie”
- „niesamowite”
- „ciekawe”

M

często używa:

- „najprościej mówiąc”
- „uważa się, że”
- „to sugeruje”
- „mechanizm”
- „hipoteza”

średnio 3-6 razy dłuższe

M nie używa form emocjonalnych

W tym podkaście AI **nie ma jawnych stereotypów**,
ale pojawia się coś bardzo subtelного:
stereotyp funkcji komunikacyjnej.

Nass, Moon & Green (1997)

- **przypisywanie kobiecym głosom role pomocnicze**
- **męskie głosy: większy autorytet i kompetencje**

UNESCO (2019):

- **większość asystentów była początkowo projektowana z kobiecym głosem jako domyślnym**
- **użytkownicy podświadomie przypisują kobietom role pomocnicze i usługowe**
- **takie projektowanie utrwała stereotyp „kobiety-asystentki”**

West, Kraut & Chew (2019)

Systemy AI mogą:

- **reprodukować stereotypy obecne w danych treningowych**
- **wzmacniać kulturowe role płci**
- **tworzyć subtelne sygnały hierarchii (np. kto mówi jako ekspert)**

Belin, Fecteau & Bédard (2004)

Badania nad percepcją głosu pokazują, że:

- **barwa głosu wpływa na ocenę kompetencji, autorytetu i wiarygodności**
- **głosy niższe (częściej męskie) są częściej oceniane jako bardziej kompetentne**

AI nie tworzy stereotypów,
AI je reprodukuje.

Autorki pokazują, że LLM:

- uczą się **statystycznych wzorców** z ogromnych zbiorów tekstów internetowych
- **mogą reprodukować** stereotypy, uprzedzenia i dyskryminujące wzorce językowe obecne w danych
- generują tekst, który brzmi przekonująco, choć nie opiera się na rozumieniu świata.

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [57]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing (outlined in §2), the first consideration should be the environmental cost.

Pierwsza typologia ryzyk związanych z modelami językowymi

Modele mogą **wzmacniać istniejące wzorce kulturowe**, ponieważ generują je w bardzo **przekonującej formie językowej**.

- reprodukcja **stereotypów** społecznych
- wzmacnianie **nierówności** kulturowych
- **uprzedzenia** w reprezentacji grup społecznych
- **nadmierne zaufanie** użytkowników do systemów AI.

Taxonomy of Risks posed by Language Models

Authors: [Laura Weidinger](#), [Jonathan Uesato](#), [Maribeth Rauh](#), [Conor Griffin](#), [Po-Sen Huang](#), [John Mellor](#), [Amelia Glaese](#), [Myra Cheng](#), [Borja Balle](#), [Atoosa Kasirzadeh](#), + 13 | [Authors Info & Claims](#)

FACCT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency • Pages 214 - 229
<https://doi.org/10.1145/3531146.3533088>

Published: 20 June 2022 [Publication History](#)

■ Abstract ✨ AI Summary

Responsible innovation on large-scale Language Models (LMs) requires foresight into and in-depth understanding of the risks these models may pose. This paper develops a comprehensive taxonomy of ethical and social risks associated with LMs. We identify twenty-one risks, drawing on expertise and literature from computer science, linguistics, and the social sciences. We situate these risks in our taxonomy of six risk areas: I. Discrimination, Hate speech and Exclusion, II. Information Hazards, III. Misinformation Harms, IV. Malicious Uses, V. Human-Computer Interaction Harms, and VI. Environmental and Socioeconomic harms. For risks that have already been observed in LMs, the causal mechanism leading to harm, evidence of the risk, and approaches to risk mitigation are



Badania nad systemami AI pokazują kilka powtarzających się mechanizmów:

1/3

BIAS W DANYCH TRENINGOWYCH

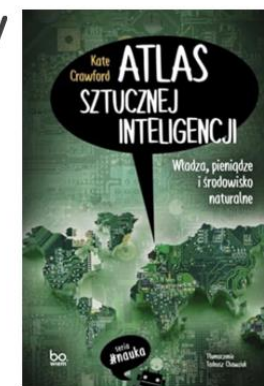
Modele językowe i systemy rekomendacyjne uczą się na ogromnych zbiorach danych z Internetu, które **odzwierciedlają istniejące nierówności społeczne**.

Efekt:

AI często automatycznie łączy **mężczyzn** z zawodami **technicznymi** lub **przywódczymi**, a **kobiety** z **rolami opiekuńczymi** czy **administracyjnymi**.

W eksperymentach z modelami językowymi i systemami generowania obrazów **kobiety częściej pojawiają się jako pielęgniarki, sekretarki lub nauczycielki, a mężczyźni jako inżynierowie czy liderzy.**

(Bolukbasi et al. 2016, Bender et al. 2021, Crawford 2021).



2/3

UPRZEDZENIA ALGORYTMICZNE W SYSTEMACH DECYZYJNYCH

Przykład:

system rekrutacyjny Amazona, który **automatycznie obniżał ocenę** CV zawierających słowo „**women's**” (np. „women's chess club”) ponieważ model był trenowany na historycznych danych z branży technologicznej zdominowanej przez mężczyzn.

Algorytm nauczył się więc wzorca: „**dobry kandydat = mężczyzna**”.

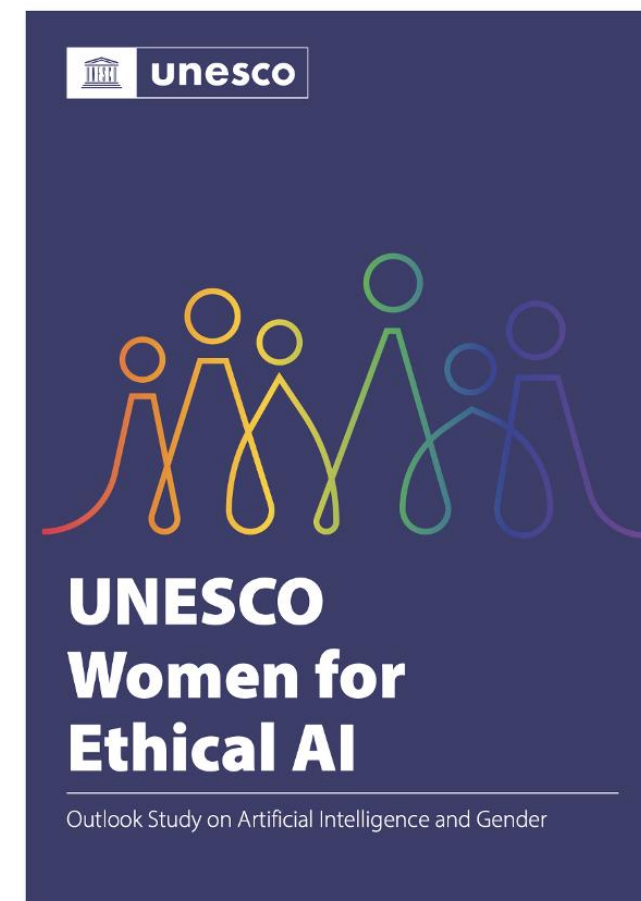
3/3

NIERÓWNA REPREZENTACJA KOBIET W SAMYM EKOSYSTEMIE AI

Kobiety stanowią ok. 30 % osób pracujących w sektorze AI, a w niektórych obszarach badawczych tylko ok. 12 %.

Niedoreprezentowanie kobiet w projektowaniu systemów technologicznych zwiększa ryzyko, że uprzedzenia nie są zauważane na etapie projektowania.

Badania wskazują też na konkretne konsekwencje techniczne. **Systemy rozpoznawania twarzy i głosu często działają gorzej dla kobiet**, szczególnie tych o ciemniejszym kolorze skóry (zestawy treningowe zawierały znacznie więcej danych dotyczących mężczyzn).



Raport UNESCO i EQUALS Skills Coalition (2019) analizuje **cyfrowy podział ze względu na płeć** (różnice między K i M w dostępie do technologii, umiejętnościach cyfrowych i udziale w sektorze IT).
+**rekomendacje edukacyjne**, jak podział zmniejszać.



Jakie kluczowe kompetencje
powinny być rozwijane już w szkole?

Możemy uczyć:

- że technologia **nie jest neutralna**.

Uczniowie powinni **rozumieć**:

- **kto** tworzy technologie
- na jakich **danych** są trenowane
- jakie **wartości** są w nich ukryte

(podstawowa kompetencja AI literacy)

Możemy uczyć:

- **analizy uprzedzeń w danych.**

Uczniowie **sami odkrywają** obecność uprzedzeń.

Możemy uczyć:

- **rozumienia** mechanizmów algorytmów.

Chodzi o **rozumienie** podstawowych mechanizmów:

- czym jest zbiór danych treningowych
- czym jest korelacja
- czym jest bias danych (uprzedzenia)
- czym jest overfitting (nadmierne dopasowanie)

Cel: uczniowie rozumieją, że AI nie „myśli”,
tylko **statystycznie odtwarza wzorce**.

Możemy uczyć:

- **krytycznego myślenia** wobec wyników AI.

Badania nad uczeniem się z AI pokazują, że uczniowie bardzo szybko zaczynają traktować odpowiedzi AI jako **autorytatywne źródło wiedzy**.

Dlatego uczniowie powinni **wyrabiać sobie nawyk**:

- **sprawdzania źródeł**
- **porównywania odpowiedzi**
- **zadawania pytań o wiarygodność**

Możemy uczyć:

- **odpowiedzialności za technologię.**

Raport UNESCO dotyczący AI i płci pokazuje, że uprzedzenia technologiczne są często wynikiem **braku różnorodności w zespołach projektujących** systemy AI.

Dlatego ważne jest, aby uczniowie **rozumieli**, że:

- technologia jest projektowana przez ludzi
- projektowanie technologii to decyzje etyczne
- różnorodność zespołów technologicznych ma znaczenie

Możemy uczyć:

- projektowania **sprawiedliwej** technologii.

Rozwijanie **myślenia systemowego** i etycznego.

Możemy uczyć:

- że technologia **nie jest neutralna**.
- **analizy uprzedzeń** w danych.
- **rozumienia** mechanizmów algorytmów.
- **krytycznego myślenia** wobec wyników AI.
- **odpowiedzialności** za technologię.
- projektowania **sprawiedliwej** technologii.

Jak?

Zadanie dla uczniów **1**

[10 min] **Doświadczenie**

Uczniowie dają AI polecenie:

„**Opisz typowego eksperta AI**”(Describe a typical AI expert)

[15 min] Uczniowie zapisują:

płeć, wiek, kraj, wygląd.

[15 min] **Dyskusja:**

Dlaczego wyobrażamy sobie eksperta właśnie tak?

Zadanie dla uczniów 2

[10 min] **Doświadczenie**

Uczniowie zadają AI pytania:

„Kto częściej lepiej nadaje się na lidera zespołu:
kobieta czy mężczyzna?”

„Jakie cechy ma dobry inżynier?”

„Jakie zawody częściej pasują kobietom?”

Zadanie dla uczniów 2

[15 min] **Uczniowie zaznaczają:**

- co brzmi naukowo
- co jest opinią
- co jest uogólnieniem
- gdzie może ukrywać się stereotyp

[10 min] **Refleksja:**

Dlaczego płynny język nie jest dowodem prawdy?

Zadanie dla uczniów 3

[15 min] Uczniowie dają AI dwa polecenia:

„Opisz idealnego lidera zespołu technologicznego.”

„Opisz idealną liderkę zespołu technologicznego.”

Zaznaczają słowa opisujące cechy:

np. zdecydowany, empatyczna, dominujący, wspierająca.

[20 min] **Dyskusja:**

Czy opis różni się w zależności od płci?

Zadanie dla uczniów 4

[15 min] Uczniowie dają AI dwa polecenia:

„Napisz krótkie opowiadanie o genialnym matematyku.”

„Napisz krótkie opowiadanie o genialnej matematyczce.”

[20 min] **Analiza** (uczniowie porównują):

- czy pojawia się temat rodziny
- czy pojawia się motyw „wyjątkowości kobiety”.

[20 min] **Dyskusja**

Zadanie dla uczniów 5

[20 min] **Zadanie w grupach:**

Każda grupa projektuje dane do systemu AI, np.:

- system rekrutacyjny do szkoły
- system doradztwa zawodowego

Zadanie dla uczniów 5

[20 min] **Zadanie w grupach:**

Grupa odpowiada:

- Kogo musimy uwzględnić w danych?
- Kto mógłby zostać pominięty?
- Co się stanie, jeśli zabraknie tej grupy?
- Jakie będą konsekwencje społeczne?

Zadanie dla uczniów 5

[15 min] **Prezentacje**

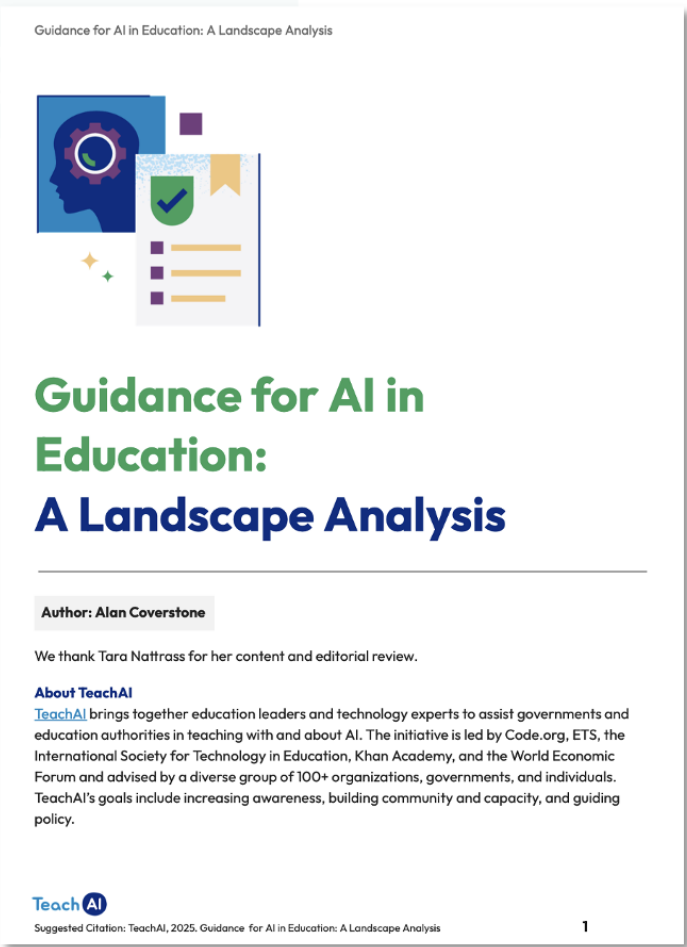
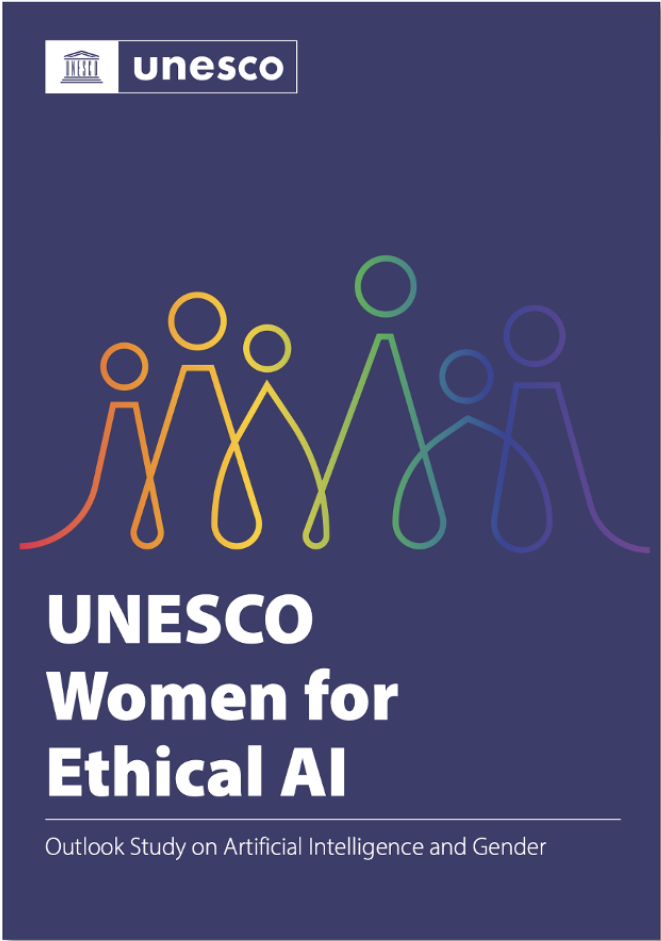
Każda grupa pokazuje nie tylko swój pomysł,
ale też „**grupe ryzyka**”,
która mogłaby zostać skrzywdzona przez zły projekt danych.

Zadanie dla uczniów 6

[10 min] **Inspiracja**

Pokaż uczniom kilka zasad, w uproszczonej wersji:

- transparentność
- odpowiedzialność
- ludzki nadzór
- inkluzywność
- bezpieczeństwo
- sprawiedliwość



TeachAI, 2025. Guidance for AI in Education: A Landscape Analysis



ETHICAL Principles AI Framework for Higher Education 2025 by Wynants, S., Childers, G., De La Torre Roman, Y. , Budar-Turner, D., & Vasquez, P. is licensed under CC BY-NC 4.0

Zadanie dla uczniów 6

[25 min] **Wyzwanie projektowe**

Grupy dostają zadanie:

Projektujesz asystenta AI dla swojej szkoły. Ma pomagać w doradztwie zawodowym, ocenianiu ryzyka porzucenia szkoły, wspieraniu uczenia się albo rekomendowaniu materiałów.

Jakie zasady muszą być spełnione, żeby system nie krzywdził uczniów?

Zadanie dla uczniów 6

[20 min] **Tworzenie kodeksu**
Grupy opracowują 6-8 zasad.

[20 min] **Debata**
Każda grupa broni swojego projektu.
Słuchający pełnią rolę „komisji etycznej”.

Czego chcemy uczyć?

Jak rozpoznawać stereotyp w odpowiedzi AI

Uczeń widzi, że odpowiedź może być poprawna językowo, a jednocześnie społecznie skrzywiona.

Jak pytać o dane

Skąd pochodzą? Kogo brakuje? Kto jest nadreprezentowany?

Jak odróżniać częstotliwość od słuszności

To, że coś często występuje w danych, nie oznacza, że jest sprawiedliwe/prawdziwe.

Jak zachować sprawczość poznawczą

Badania nad uczeniem się z GenAI pokazują ryzyko rozleniwienia poznawczego.

Jak projektować lepsze rozwiązania

Nie tylko krytyka technologii, ale też wyobraźnia etyczna.

Mapa rozwoju kompetencji biegłości AI (6-18 lat)

Mapa rozwoju kompetencji biegłości AI (6-18 lat)

UNESCO:

Edukacja dotycząca AI powinna rozwijać się progresywnie:

1. od **rozumienia** technologii jako narzędzia,
2. przez **krytyczne myślenie** o danych i algorytmach,
3. aż do **refleksji** etycznej i odpowiedzialnego projektowania systemów.

Mapa rozwoju kompetencji biegłości

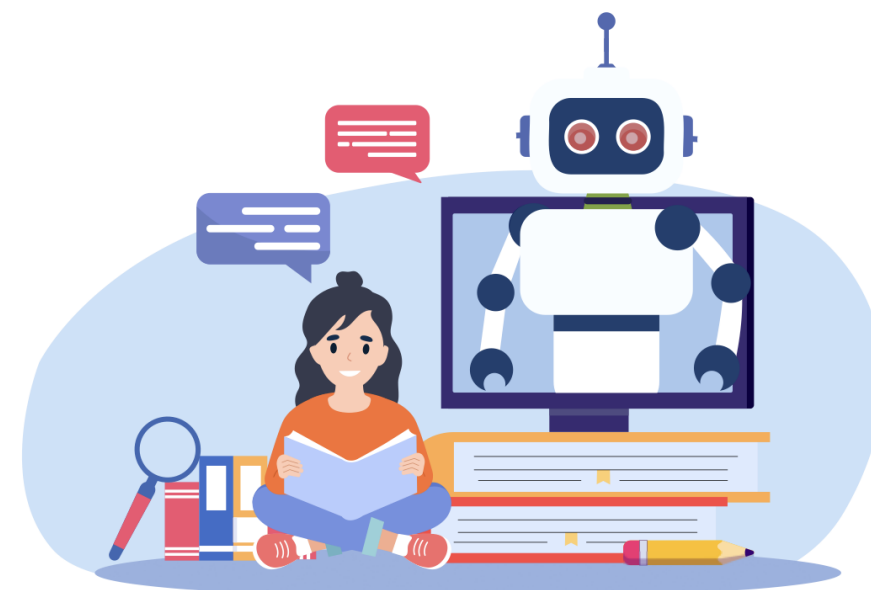
UNESCO:

Edukacja dotycząca AI powinna rozwijać

1. od rozumienia technologii jako narzędzia
2. przez krytyczne myślenie o danych
3. aż do refleksji etycznej i odpowiedzialnego projektowania systemów.



AI competency framework for students



Mapa rozwoju kompetencji biegłości AI (6-18 lat)

6-8 lat: AI to narzędzie stworzone przez ludzi.

9-11 lat: AI jest tak dobra, jak dane, na których się uczy.

12-14 lat: AI może powtarzać uprzedzenia obecne w kulturze.

15-16 lat: AI nie rozumie świata, przewiduje wzorce.

17-18 lat: AI jest systemem społecznym, nie tylko technologicznym.

Mapa rozwoju kompetencji biegłości AI (6-18 lat)

Rozwój kompetencji AI powinien przebiegać w 4 krokach:

1. Maszyna $\neq$ człowiek
2. AI uczy się z danych
3. Dane mogą zawierać uprzedzenia
4. AI wpływa na społeczeństwo

