

EDUxAI

Przewodnik dla
osób wspierających
młodych ludzi
w uczeniu się
ze wsparciem AI

dla nauczycieli

Joanna Mytnik

 TL;DR

Pobierz PDF

ISBN: 978-83-968391-5-2

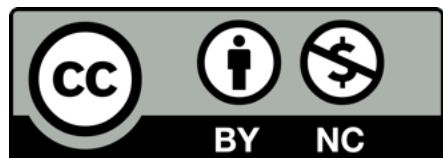
Redakcja, projekt okładki, oprawa graficzna: Joanna Mytnik

Centrum Nowoczesnej Edukacji Politechniki Gdańskiej
Gdańsk 2025

Ver. 1.0 [31 grudnia 2025 r.]

Niniejszy poradnik jest zbiorem wybranych materiałów z moich wykładów konferencyjnych. Nie stanowi kompendium wiedzy na temat generatywnej sztucznej inteligencji. Skupiłam się na wybranych zagadnieniach związanych z wykorzystaniem sztucznej inteligencji w nauczaniu. W niniejszej publikacji skrót AI odnosi się do generatywnej sztucznej inteligencji (został zastosowany dla uproszczenia przekazu).

Wykonane w app.genially.com.



CC-BY-NC 4.0 (uznanie autorstwa, cel niekomercyjny). [?](#)

Spis treści

CZĘŚĆ I.

KRYZYS UNIWERSYTETU W EPOCE POST-AI

Zmiana paradygmatu

Studenci i narzędzia AI

Produkty „edukacyjne” i interesy rynku

CZĘŚĆ II.

CIEMNA STRONA AI

Zagrożenia dla procesów poznawczych

Pochlebczość i ugodowość modeli językowych

AI companion i relacje z technologią

Zagrożenia informacyjne

Platformy demaskujące dezinformację

Budowanie odporności psychicznej

CZĘŚĆ III.

PISANIE HYBRYDOWE I KRYZYS OCENIANIA

Pisanie hybrydowe

Strategia Human-AI-Human

6 zasad postplagiatu (Sarah Eaton)

Jak wesprzeć studentów w czytaniu, pisaniu, refleksji

Jak AI wspiera personalizację uczenia się

Cechy tekstu generowanego przez AI

Ocenianie w epoce post-AI

Zmiana warsztatu dydaktycznego

Zaktualizuj swój sylabus

Wskazówki wykorzystania AI w zadaniu (skala AIAS)

Detektory AI nie rozwiązują problemu

DALSZE WSPARCIE

TL;DR

Uważaj na produkty "edukacyjne". Bądź sceptyczna/y. Zadawaj sobie pytania: czy to narzędzie rzeczywiście wspiera samodzielne myślenie studentów, proces uczenia się?

Następuje zmiana modelu uczenia się, pisanie hybrydowe staje się normą. Definicja plagiatu już nie wystarcza, definicja autorstwa rozmywa się, rośnie kryzys zaufania do źródła wiedzy.

Wspieraj biegłość AI studentów: w każdym zadaniu określ czy i jeśli tak, to na którym etapie student może użyć narzędzi AI i w jakim celu (AI log).

Wypracuj AI-odporne sposoby oceniania. Wspieraj studentów w nabywaniu biegłości pisania, czytania, refleksji i krytycznego myślenia.

Największym wyzwaniem dla nas, nauczycieli,
jest **uświadomienie sobie** skali i charakteru zmian
zachodzących za sprawą AI
oraz **refleksyjne projektowanie scenariuszy przyszłości**,
które pozwalają dostrzegać ich konsekwencje
dla procesu uczenia się i rozwoju człowieka.

I przeciwdziałać im zawczasu.

KRYZYS UNIWERSYTETU W EPOCE **POST-AI**

Zmiana paradygmatu

Od stuleci:

- informacja była dobrem rzadkim, trudnym do zdobycia
- uczelnia była strażnikiem prawdy i autorytetu
- uczenie się polegało na przyswajaniu tego, co powszechnie uznane było za wiarygodne

Dziś:

- **informacja jest łatwo dostępna**
- **definicja autorstwa się rozmywa**
- **granica zaufania do źródła wiedzy przesuwana się**

Technologia AI **zmienia sposób prowadzenia badań oraz nauczanie** i uczenie się, dlatego pilnie potrzebujemy **zdecydowanych działań** w tym zakresie.

Po 3 latach nadal brak **spójnej i strategicznej wizji** oraz **stanowczych działań** dotyczących **AI w edukacji wyższej**.

Największym wyzwaniem dla szkolnictwa wyższego jest instytucjonalna inercja w kontekście technologii AI.

"**Uczelnie** ryzykują utratę znaczenia, bo **wciąż nie angażują** się poważnie w sztuczną inteligencję.

Trudno wskazać inny sektor, który tak wyraźnie **przegapił strategiczne wykorzystanie potencjału nowej technologii.**"

Ian Richardson

THE Times Higher Education

[Home](#)[News](#)[Rankings](#)[Business](#)[Jobs](#)

Universities risk irrelevance by failing to engage fully with AI

It is difficult to think of another sector that has so dismally failed to strategically engage with the transformative potential of IT, says Ian Richardson

Published on November 11, 2025

Last updated November 11, 2025

[Ian Richardson](#)

Universities perform a function in our societies that they alone can perform. And, given the transcendental value of that role, they should be allowed to perform it without interference, oversight or criticism.



Source: sshepard/Getty Images

Cloaked in Enlightenment principles, such an uncompromising and self-regarding belief has, in recent decades, seen the higher education sector drift – detached, cautious and often unwilling to reimagine its purpose in the face of changing societal expectations, even amid the steady creep of marketisation.

Despite a growing acceptance that the current model is unsustainable, few institutions have demonstrated an inclination to challenge antiquated and conflicted governance structures, risk-averse bureaucracies, opaque decision-making cultures, dated curricula and pedagogies, questionable contributions and standards of quality, and ever-mounting costs.

W świecie współtworzonym przez człowieka i AI granice **autorstwa, wkładu i oryginalności** stają się płynne.

AI nie jest jedynie narzędziem. AI zmienia ekosystem wiedzy, sposoby tworzenia tekstów, modele uczenia się i to, co uznajemy za „dobrą pracę intelektualną”.

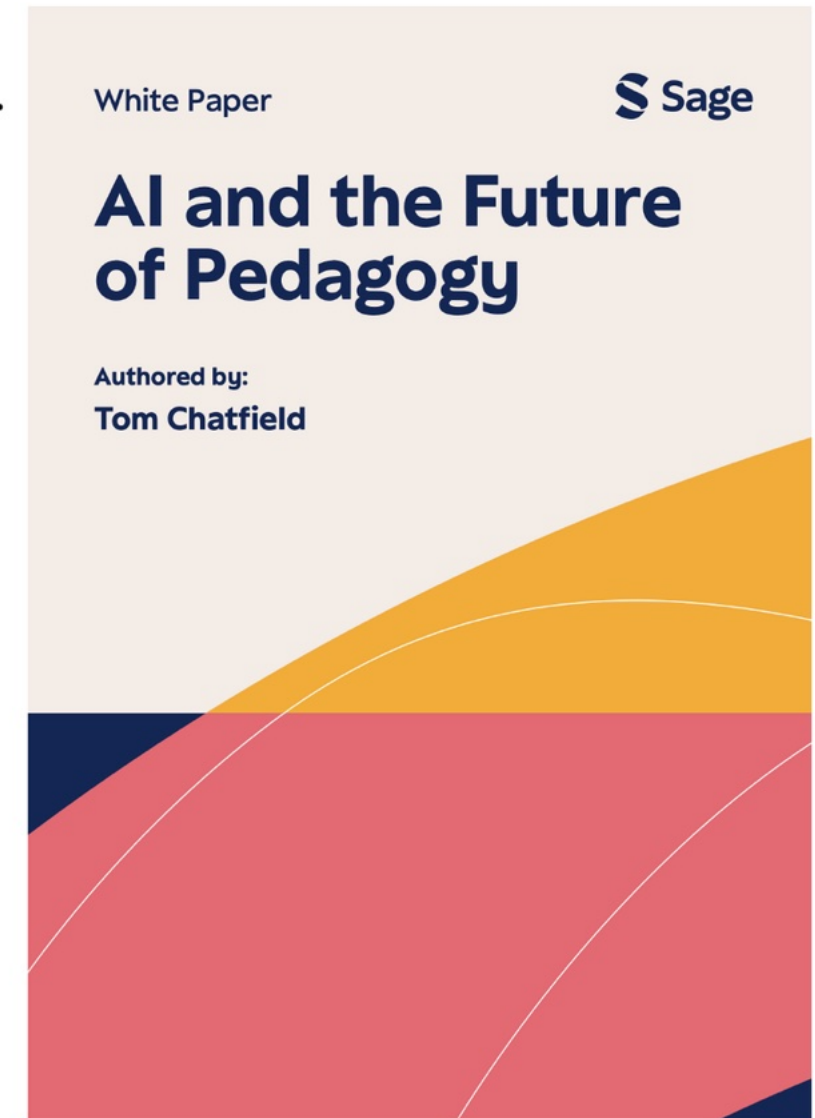
Największym ryzykiem **nie jest** technologia, ale utrata **refleksyjności**.

Nie chodzi o dodanie AI do starych praktyk, ale o:

- redefinicję oceniania,
- przeprojektowanie zadań/instrukcji,
- stworzenie pól do refleksji,

O zrozumienie, jak **AI przekształca procesy uczenia się**.

Brak rozmowy o AI prowadzi do ukrywania użycia narzędzi, wypierania ich roli, erozji zaufania, „postplagiatu”.



„Niebezpieczeństwem nie jest plagiat, a **erozja zaufania publicznego** do uniwersytetów jako gwarantów wiarygodności. Gdy zaufanie się załamuje, **rozpada się fundament społecznej umowy.**”

W jaki sposób uczelnie wyższe na całym świecie muszą przededefiniować swoje funkcje i strategie w obliczu technologii AI?

Trzy główne siły napędowe:

1. transformację badań i nauczania,
2. konieczność elastycznych ścieżek kształcenia,
3. pogłębienie współpracy między uczelniami a przemysłem.

Uczelnie stoją przed wyborem:

będą ewoluować lub zostaną zmarginalizowane.

James Yoonil Auh, professor at KyungHee Cyber University Chair of the Department of Computer and Information Communications Engineering.

University World News 
THE GLOBAL WINDOW ON HIGHER EDUCATION

GLOBAL

The end of AI and the future of higher education

James Yoonil Auh 01 October 2025

 Share

 Post

 Udostępnij 40

For more than a decade, universities treated artificial intelligence as a procurement decision or a narrow policy debate. Should students be allowed to use ChatGPT on essays? Should professors rely on automated grading? Should institutions buy yet another platform promising machine-learning efficiency?

Those debates are already over. The question is no longer whether higher education should adopt AI, but what comes after adoption, in a world where AI has ceased to exist as a discrete category.

We now live in what I call the atmosphere of cognition: not the disappearance of AI, but its absorption into the invisible architectures of institutional life. Like Wi-Fi, AI is no longer a tool at the margins but the infrastructure of thought.

From tool to atmosphere

Algorithms now shape admissions. Predictive models determine financial aid. Recommendation engines curate research. Plagiarism detectors and manuscript filters run silently in the background. To speak of 'using AI' in 2025 is like debating whether universities should install electricity.

This is not simply a technical evolution. It is civilisational. The printing press multiplied texts, but students still thought alone. The internet digitised knowledge, but students still wrote in their own words. MOOCs disrupted delivery but not learning itself.

5 zasad dla uniwersytetu post-AI

Integralność poznawcza: studenci muszą zrozumieć nie tylko, jak pisać lub obliczać, ale także, w jaki sposób algorytmy subtelnie kształtują sposób, w jaki uczą się, myślą.

Suwerenność danych: instytucje muszą chronić własność i zarządzanie tymi danymi, aby zachować wolność akademicką i zaufanie.

Odpowiedzialność ekologiczna: systemy AI zależą od ogromnej ilości energii i wody. Uniwersytety nie mogą wiarygodnie uczyć zrównoważonego rozwoju ignorując koszty środowiskowe infrastruktury cyfrowej.

Uczenie się relacyjne: dostarczanie treści może być replikowane przez maszyny. To, czego nie można zautomatyzować, to mentoring, zaufanie i dialog. Relacje międzyludzkie pozostają w centrum transformacyjnego uczenia się.

Sprawiedliwość epistemiczna: uniwersytety muszą aktywnie bronić różnorodności intelektualnej i zapewnić, że lokalne, rdzenne i alternatywne perspektywy nie zostaną wymazane przez standaryzację.

GLOBAL

The end of AI and the future of higher education

James Yoonil Auh 01 October 2025

 Share

 Post

 Udostępnij 40

For more than a decade, universities treated artificial intelligence as a procurement decision or a narrow policy debate. Should students be allowed to use ChatGPT on essays? Should professors rely on automated grading? Should institutions buy yet another platform promising machine-learning efficiency?

Those debates are already over. The question is no longer whether higher education should adopt AI, but what comes after adoption, in a world where AI has ceased to exist as a discrete category.

We now live in what I call the atmosphere of cognition: not the disappearance of AI, but its absorption into the invisible architectures of institutional life. Like Wi-Fi, AI is no longer a tool at the margins but the infrastructure of thought.

From tool to atmosphere

Algorithms now shape admissions. Predictive models determine financial aid. Recommendation engines curate research. Plagiarism detectors and manuscript filters run silently in the background. To speak of 'using AI' in 2025 is like debating whether universities should install electricity.

This is not simply a technical evolution. It is civilisational. The printing press multiplied texts, but students still thought alone. The internet digitised knowledge, but students still wrote in their own words. MOOCs disrupted delivery but not learning itself.

7 wątpliwych założeń na temat AI w edukacji

” **Większa personalizacja = efektywniejsze uczenie się.** Modne hasło dekady, ale czym właściwie jest "personalizacja"?

AI zastąpi nauczycieli. Tyle tylko, że ktoś nadal musi interpretować efekty pracy AI, zarządzać interakcjami między uczniami, "sprzątać" po halucynacjach, rozmawiać o wpływie na środowisko i problemach etycznych. Czy to nie więcej pracy "w przebraniu" odciążenia?

Człowiek + AI oznacza lepsze myślenie. Tylko dla niektórych. Istnieje różnica między tworzeniem elegancko brzmiącego tekstu, a umiejętnością krytycznego myślenia.

Musimy uczyć umiejętności korzystania z AI. I tak, i nie.

AI rozwiązuje problem szybkiej informacji zwrotnej. To proces szybki, skalowalny i całkowicie oderwany od kontekstu. Dobra informacja zwrotna wymaga zaufania i profesjonalnej wiedzy. AI może to naśladować, ale to powierzchowne działanie.

AI zdemokratyzuje edukację. Tylko jeśli zignorujesz, kto kontroluje narzędzia i kogo stać na wersję premium.

AI będzie dużo lepsza za 6 miesięcy. Może. Ale musimy pracować z tym, co jest tu i teraz. Czekanie na następną wersję nie rozwiąże dzisiejszych problemów.

Za: Jeppe Klitgaard Stricker (zmienione)

STUDENCI I NARZĘDZIA AI

”

W edukacji utożsamiamy ChatGPT z oszukiwaniem. To sprawia, że nie dostrzegamy tego, jak **technologia po cichu przekształca nasz świat.**

Musimy zacząć traktować generatywną sztuczną inteligencję jako **technologię kulturową, która zmienia społeczeństwo.**

Niebezpieczeństwa związane z siedzeniem z założonymi rękami i brakiem dyskusji ze studentami na temat tego, jak ta nowa era technologiczna kształtuje nasz świat, są zbyt wielkie, abyśmy mogli milczeć.

Marc Watkins

Uczelnie wysyłają niejasne i niespójne sygnały.

Studenci czują, że uczelnie narzucają surowe zasady, jednocześnie nie znają ich, polegają na plotkach i domysłach („detektory” AI).

W efekcie **boją się używać** AI jednocześnie czując, że **muszą jej używać**.
Próbują balansować między efektywnością a ryzykiem.

Studenci uważają, że ich praca powinna pochodzić od nich i to jest ważniejsze niż reguły formalne.

„**Własność pracy**” oznacza dla studentów:

- że pomysły są ich,
- że proces jest ich,
- że rozumieją to, co oddają,
- że nie czują się „zastąpieni”.

HIGHER EDUCATION RESEARCH & DEVELOPMENT
<https://doi.org/10.1080/07294360.2025.2580616>

 **Routledge**
Taylor & Francis Group

 OPEN ACCESS  Check for updates

Time, emotions and moral judgements: how university students position GenAI within their study

Margaret Bearman ^a, T. Fawns^b, T. Corbin^a, M. Henderson^c, Y. Liang^d, G. Oberg^d, J. Walton^a and K. E. Matthews^d

^aCentre for Research in Assessment and Digital Learning (CRADLE), Deakin University, Melbourne, Australia;

^bMonash Education Academy, Monash University, Melbourne, Australia; ^cFaculty of Education, Monash University, Melbourne, Australia; ^dOffice of the Deputy Vice-Chancellor (Academic), University of Queensland, Brisbane, Australia

ABSTRACT

The emergence of Generative AI (GenAI) in higher education has prompted considerable discussion within the research community. Despite their centrality, students' perspectives remain underexplored. We investigated how students position GenAI in relation to their studies, conducting focus groups with 79 students across four Australian universities. Taking a sociotechnical stance and employing reflexive thematic analysis, we identified three primary themes: (1) studying with and without GenAI; (2) mixing messages and assumptions; and (3) 'coming from me': self-trust and resistance to dependency. Crossing these themes were axial threads of time, emotions, and moral judgement. Our findings illuminate a complex, dynamic and uncertain landscape of relationships in which students prioritise their developing values and moral positions over institutional messaging. Amongst diverse and changing practices, students generally conveyed a sense that their educational work should come from, and be owned by, themselves. This concept of ownership could productively inform academic integrity discussions. Institutions may need to move beyond rules and restrictions, to foster environments where students can develop nuanced relationships with GenAI whilst maintaining academic autonomy. This research demonstrates the value of centring student voices in shaping scholarly practice and policy, particularly in contexts of significant technological change.

ARTICLE HISTORY

Received 4 November 2024
Accepted 12 October 2025

KEYWORDS

Artificial intelligence;
generative artificial
intelligence; sociotechnical;
qualitative methods;
academic integrity; higher
education

Studenci obawiają się uzależnienia od AI.

Wielu mówi o:

- **lęku** przed utratą umiejętności,
- **obawie**, że AI „rozleniwi”,
- **strachu**, że w egzaminie zostaną bez narzędzi i „to się wyda”.

Studenci chcą przewodnictwa, ale nie kontroli.

Badanie pokazuje, że **studenci potrzebują**:

- jasnych zasad,
- wskazówek, jak uczyć się z AI,
- zrozumieć, czego uczelnia od nich oczekuje.

Uczelnie są postrzegane jako **spóźnione, niepewne, często milczące.**

HIGHER EDUCATION RESEARCH & DEVELOPMENT
<https://doi.org/10.1080/07294360.2025.2580616>



OPEN ACCESS

Time, emotions and moral judgements: how university students position GenAI within their study

Margaret Bearman ^a, T. Fawns^b, T. Corbin^a, M. Henderson^c, Y. Liang^d, G. Oberg^d, J. Walton^a and K. E. Matthews^d

^aCentre for Research in Assessment and Digital Learning (CRADLE), Deakin University, Melbourne, Australia; ^bMonash Education Academy, Monash University, Melbourne, Australia; ^cFaculty of Education, Monash University, Melbourne, Australia; ^dOffice of the Deputy Vice-Chancellor (Academic), University of Queensland, Brisbane, Australia

ABSTRACT

The emergence of Generative AI (GenAI) in higher education has prompted considerable discussion within the research community. Despite their centrality, students' perspectives remain underexplored. We investigated how students position GenAI in relation to their studies, conducting focus groups with 79 students across four Australian universities. Taking a sociotechnical stance and employing reflexive thematic analysis, we identified three primary themes: (1) studying with and without GenAI; (2) mixing messages and assumptions; and (3) 'coming from me': self-trust and resistance to dependency. Crossing these themes were axial threads of time, emotions, and moral judgement. Our findings illuminate a complex, dynamic and uncertain landscape of relationships in which students prioritise their developing values and moral positions over institutional messaging. Amongst diverse and changing practices, students generally conveyed a sense that their educational work should come from, and be owned by, themselves. This concept of ownership could productively inform academic integrity discussions. Institutions may need to move beyond rules and restrictions, to foster environments where students can develop nuanced relationships with GenAI whilst maintaining academic autonomy. This research demonstrates the value of centring student voices in shaping scholarly practice and policy, particularly in contexts of significant technological change.

ARTICLE HISTORY

Received 4 November 2024
Accepted 12 October 2025

KEYWORDS

Artificial intelligence;
generative artificial
intelligence; sociotechnical;
qualitative methods;
academic integrity; higher
education

”

"Przez ponad 25 lat uczyłem systemów informatycznych tysiące studentów w University College Cork i **NIGDY nie pracowałem z pokoleniem uczniów bardziej intelektualnie, etycznie i osobiście zaangażowanym w technologię i jej wpływ na uczenie się**, pracę, wiedzę, kreatywność, innowacje, społeczeństwo i na to, co to znaczy być człowiekiem w sensie ludzkim, niż teraz.

Nasi studenci aktywnie angażują się w sztuczną inteligencję, jej ograniczenia, wpływ, skutki i koszty, i robią to z ciekawością, kreatywnością, współczuciem i krytyczną refleksją. Możemy się od nich wiele nauczyć.

Przestań zgadywać. Porozmawiaj ze swoimi studentami."

Joseph Feller

PRODUKTY I USŁUGI

"EDUKACYJNE"

Edukacja to dochodowy sektor.

Uruchom krytyczne myślenie, gdy widzisz usługi lub produkty
"wspierające uczenie się".

Stawiaj pytania: Kto kontroluje narzędzia "edukacyjne"? Kto i co zyskuje? Jaka metodyka nauczania stoi za daną usługą? Czy wszystkich moich studentów stać na wersje Pro?

Bądź sceptyczny/a.

Rozmawiaj o tym z uczniami i studentami.

Jeśli firmy technologiczne mogły wykorzystać **wyniki badań** do tworzenia uzależniającej technologii (algorytmy mediów społecznościowych), > dlaczego nie mogą tego zrobić dla **wzmocnienia procesu uczenia się i wsparcia zdrowia psychicznego**?
Domagajmy się tego od EdTech'u.

Foteliki samochodowe i zabawki wymagają atestów, a co z produktami edukacyjnymi?

Edukacja, nie tylko zysk



1. Potrzebujemy, by firmy współpracujące ze szkołami i uczniami działały jako **Benefit Corporation** (podejmujące decyzje o zmniejszaniu zysków na rzecz dobra społecznego).
2. Wymagajmy **bezpieczeństwa produktu** (jak foteliki samochodowe czy zabawki).
3. Oczekujmy **definicji umiejętności korzystania z AI**, aby uwzględnić etyczne wykorzystanie technologii.
4. Umieśćmy **nauczycieli w centrum**, powinni być zaangażowani w projektowanie wykorzystania AI w edukacji.
5. Zwiększmy **prawdziwe zaangażowanie** uczniów i studentów.

Rebecca Winthrop, director of the Center for Universal Education at Brookings and an adjunct professor at Georgetown University

”

Firmy zajmujące się generatywną sztuczną inteligencją budują nową infrastrukturę poznania, a my wynajmujemy w nim powierzchnię. (...)

Kto decyduje o tym, jak wygląda myślenie? Za każdym razem, gdy uczeń używa ChatGPT do "zrozumienia" koncepcji, zleca tę decyzję danym szkoleniowym, zasadom treści i celom optymalizacyjnym OpenAI. Za każdym razem, **gdy instytucja wdraża narzędzie** do pisania oparte na sztucznej inteligencji, **akceptuje czyjaś definicję "dobrego pisania"**. (...)

Tracimy suwerenność epistemologiczną, prawo do decydowania o tym, **co liczy się jako wiedza, uczenie się i myślenie w kontekście edukacyjnym**. Firmy zajmujące się sztuczną inteligencją to nie tylko narzędzia do tworzenia. Budują nowe standardy myślenia. Każdego dnia, kiedy z nich korzystamy, **uczymy uczniów, że myślenie odbywa się wewnątrz czyjegoś systemu**, zgodnie z **czyimiś zasadami**, zoptymalizowanymi pod **czyjeś cele**."

Jeppe Klitgaard Stricker

CIEMNA STRONA AI

1. Zagrożenia dla procesów poznawczych
2. Pochlebność i ugodowość LLM
3. AI companion (towarzysze AI)
4. Zagrożenia informacyjne

ZAGROŻENIA DLA PROCESÓW POZNAWCZYCH

Kalkulator

Dlaczego uczymy dzieci mnożenia bez kalkulatora?

Chodzi o rozwijanie umiejętności myślenia matematycznego.

Iluzja mistrzostwa

Wyniki badań coraz częściej sugerują, że modele językowe AI są nie tylko nieoptymalne dla procesu uczenia się, ale mogą mieć **negatywny wpływ na rozwój** wiedzy i umiejętności. Korzystanie z AI może prowadzić do **nadmiernego polegania** na tych narzędziach i w efekcie **osłabiać** zdolność **samodzielnego** rozwiązywania problemów i rozwoju umiejętności zwłaszcza u osób początkujących.

Stosowanie narzędzi AI przez **osoby o niskim** poziomie krytycznego myślenia, **osłabia** proces uczenia się podstawowych umiejętności, które są niezbędne do myślenia na wyższym poziomie.

Istnieje związek między niskim poziomem krytycznego myślenia i bezrefleksyjnym poleganiu na efektach pracy narzędzi AI: **osoby z najniższym poziomem** umiejętności krytycznego myślenia są najczęściej **najbardziej zależne od AI**.

Popatrzmy, co mówią wyniki badań



- dostęp do chatGPT znacznie **podnosi wydajność** osób uczących się (48% w przypadku chatGPT wersja podstawowa i 127% dla GPT-tutor)
- **po odebraniu tego dostępu: osoby korzystające z chatGPT osiągają gorsze wyniki niż osoby nigdy nie korzystające ze wsparcia AI (17%)**
- skutki te są **mniej negatywne** w przypadku pracy z GPT-tutor (dzięki obecnym w nim zabezpieczeniom)
- wyniki wskazują na możliwy **negatywny** wpływ nadmiernego polegania na AI na trwałość uczenia się, szczególnie u osób o niskim poziomie samoregulacji i krytycznego myślenia.

Generative AI Can Harm Learning

The Wharton School Research Paper

59 Pages • Posted: 18 Jul 2024

[Hamsa Bastani](#)

University of Pennsylvania - The Wharton School

[Osbert Bastani](#)

University of Pennsylvania - Department of Computer and Information Science

[Alp Sungu](#)

University of Pennsylvania

[Haosen Ge](#)

University of Pennsylvania - The Wharton School

[Özge Kabakcı](#)

Budapest British International School

[Rei Mariman](#)

Independent; Independent

Date Written: July 15, 2024

Abstract

Generative artificial intelligence (AI) is poised to revolutionize how humans work, and has already demonstrated promise in significantly improving human productivity. However, a key remaining question is how generative AI affects *learning*, namely, how humans acquire new skills as they perform tasks. This kind of skill learning is critical to long-term productivity gains, especially in domains where generative AI is fallible and human experts must check its outputs. We study the impact of generative AI, specifically OpenAI's GPT-4, on human learning in the context of math classes at a high school. In a field experiment involving nearly a thousand students, we have deployed and evaluated two GPT based tutors, one that mimics a standard ChatGPT interface (called GPT Base) and one with prompts designed to safeguard learning (called GPT Tutor). These tutors comprise about 15% of the curriculum in each of three grades. Consistent with prior work, our results show that access to GPT-4 significantly improves performance (48% improvement for GPT Base and 127% for GPT Tutor). However, we additionally find that when access is subsequently taken away, students actually perform worse than those who never had access (17% reduction for GPT Base). That is, access to GPT-4 can harm educational outcomes. These negative learning effects are largely mitigated by the safeguards included in GPT Tutor. Our results suggest that students attempt to use GPT-4 as a "crutch"

Neuronowe reprezentacje wiedzy tworzą się dzięki długotrwałemu wysiłkowi.

**Hebbowska reguła
plastyczności
synaptycznej**

“neurons that fire together
wire together”
regularna aktywacja sieci
neuralnej wzmacnia
połączenia synaptyczne

**Konsolidacja
długoterminowa**

trwałe ślady pamięciowe
wymagają powtarzania,
uwagi i snu (przenoszenie
z hipokampa do kory)

**Zjawisko “effortful
processing”**

głębokie uczenie się
wymaga aktywnego,
świadomego przetwarzania
informacji

**Zasada
“spaced repetition”**

“spacing learning”
powtórki w coraz większych
odstępach czasu

Tworzenie śladów pamięciowych
powstaje dzięki
długotrwałemu wysiłkowi.

To jak **wydeptywanie ścieżki** wśród
zarośli, której na początku nie ma:
im więcej razy przechodzisz tą samą
trasą, tym staje się to łatwiejsze.
Za każdym razem potrzeba mniej
wysiłku na przejście.



Uczenie się = wykrywanie różnicy między oczekiwaniem a wynikiem

Układ dopaminowy mózgu koduje **błędy przewidywań** (prediction error).

- Zaskoczenie → wzrost dopaminy → „To jest ważne, zapamiętaj to!”
- Brak błędu → brak sygnału → brak uczenia się

Jeśli AI podaje idealne odpowiedzi, osoba ucząca się nigdy nie doświadcza tego „zaskoczenia”.

Brak oporu = brak wzmocnienia = brak trwałego uczenia się.

The Memory Paradox: Why Our Brains Need Knowledge in an Age of AI

50 Pages • Posted: 14 May 2025 • Last revised: 19 Jun 2025

[Barbara Oakley](#)

Oakland University; The New Zealand Initiative

[Michael Johnston](#)

affiliation not provided to SSRN

[Kenzen Chen](#)

affiliation not provided to SSRN

[Eulho Jung](#)

Uniformed Services University of the Health Sciences

[Terrence Sejnowski](#)

Salk Institute for Biological Studies

Date Written: May 11, 2025

Abstract

In an era of generative AI and ubiquitous digital tools, human memory faces a paradox: the more we offload knowledge to external aids, the less we exercise and develop our own cognitive capacities. This chapter offers the first neuroscience-based explanation for the observed reversal of the Flynn Effect—the recent decline in IQ scores in developed countries—linking this downturn to shifts in educational practices and the rise of cognitive offloading via AI and digital tools. Drawing on insights from neuroscience, cognitive psychology, and learning theory, we explain how underuse of the brain's declarative and procedural memory systems undermines reasoning, impedes learning, and diminishes productivity. We critique contemporary pedagogical models that downplay memorization and basic knowledge, showing how these trends erode long-term fluency and mental flexibility. Finally, we outline policy implications for education, workforce development, and the responsible integration of AI, advocating strategies that harness technology as a complement to – rather than a replacement for – robust human knowledge.

Keywords: cognitive offloading, memory, neuroscience of learning, declarative memory, procedural memory, generative AI, Flynn Effect, education reform, schemata, digital tools, cognitive load, cognitive architecture.

Poprzez **wyszukiwanie informacji i łączenie ich**, budujesz swoją wiedzę. Dzięki tej pracy możesz korzystać z niej "automatycznie", bo tworzą się schematy: **ramy mentalne, które wspierają efektywne myślenie**.

Natomiast

poleganie na zewnętrznych narzędziach omija ten proces, prowadząc do pozornego rozumienia.

W efekcie

pomimo że masz dostęp do informacji, masz **trudności** z płynnym myśleniem, dostrzeganiem błędów lub procesem kreatywnym.

Narzędzia AI mogą usprawnić proces uczenia się, ale **nie powinny zastępować procesu rozumienia i internalizacji wiedzy**.

Paradoks ery cyfrowej: podczas gdy informacja jest dostępna, **głębokie uczenie się** nadal zależy od **wiedzy, którą nosimy w sobie**.

za: Jonathan Boymal (zmienione)

Joanna Mytnik | EDUxAI dla nauczycieli | 2025 | CC-BY-NC 4.0

Paradoks ery cyfrowej:
podczas gdy informacje są łatwo dostępne,
głębokie uczenie się nadal zależy od
wiedzy, którą nosimy w sobie.

Kiedy angażujesz się w proces tworzenia/uczenia się, sukcesywnie **budujesz rusztowanie, które następnie stopniowo zanika wraz ze wzrostem kompetencji.**

Ale **kiedy pozwalasz narzędziom AI** budować to rusztowanie (np. gdy jeszcze jesteś nowicjuszem w danej dziedzinie), usuwasz tę walkę, która umożliwia proces uczenia się i rozwoju.

W efekcie zadania są wykonywane, ale akt uczenia się cicho znika. Cicho, bo nie zauważasz tego, widząc efekty wydaje Ci się, że się rozwijasz. To iluzja mistrzostwa (złudne poczucie rozwoju).

A tylko dzięki temu wysiłkowi możesz budować swoją wiedzę i umiejętności.

Podejmujesz ryzyko, jeśli nigdy naprawdę się nie angażujesz.

Czym ryzykujesz?

Rusztowanie zbudowane przez narzędzie AI uporczywie po cichu zastępuje wysiłek, który sprawia, że proces uczenia się zachodzi.

Za: Hans Visser (zmienione)



Uczenie się to tworzenie nowych połączeń w mózgu.

Powstają nowe połączenia między neuronami (komórkami nerwowymi). Nowe informacje są kodowane przede wszystkim w hipokampie i korze nowej. Świeżo powstałe połączenia są na początku słabe.



Wysiłek i frustracja to sygnał, że te drogi właśnie się wykuwają.

Nowe połączenia, początkowo słabiutkie, z czasem - dzięki wysiłkowi intelektualnemu, dzięki powtarzaniu, przywoływaniu informacji z pamięci - ulegają wzmocnieniu. Moment trudności i „oporu” to efekt tego, że świeżo powstałe ścieżki neuronalne są jeszcze niestabilne. Frustracja, która pojawia się przy próbie rozwiązania problemu/pisania pracy, wiąże się z większą aktywacją kory przedczołowej, co sprzyja głębszemu przetwarzaniu informacji i budowaniu bardziej trwałych reprezentacji w mózgu.



Bez tej walki nie powstaje trwała pamięć, a nowe informacje "wyparują" (ścieżki zanikają).

Ślady pamięciowe, które nie zostają aktywnie przywołane, ulegają wygaszaniu: połączenia między neuronami słabną, a informacja nie przechodzi procesu konsolidacji w pamięci długotrwałej.

Dług poznawczy (cognitive debt)

Konsekwencja rezygnacji z wysiłku:

rozwiązanie się pojawia, ale ścieżka myślenia nie została przetarta.

To jak czytanie streszczenia zamiast książki: wydaje się, że coś wiemy, ale nie mamy na czym budować dalej.

Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task^Δ

Nataliya Kosmyrna ¹	Eugene Hauptmann	Ye Tong Yuan	Jessica Situ
MIT Media Lab	MIT	Wellesley College	MIT
Cambridge, MA	Cambridge, MA	Wellesley, MA	Cambridge, MA

Eugene Hauptmann
MIT
Cambridge, MA

Ye Tong Yuan
Wellesley College
Wellesley, MA

Jessica Situ
MIT
Cambridge, MA

Xian-Hao Liao
Mass. College of Art
and Design (MassArt)
Boston, MA

Ashly Vivian Beresnitzky
MIT
Cambridge, MA

Iris Braunstein
MIT
Cambridge, MA

Pattie Maes
MIT Media Lab
Cambridge, MA

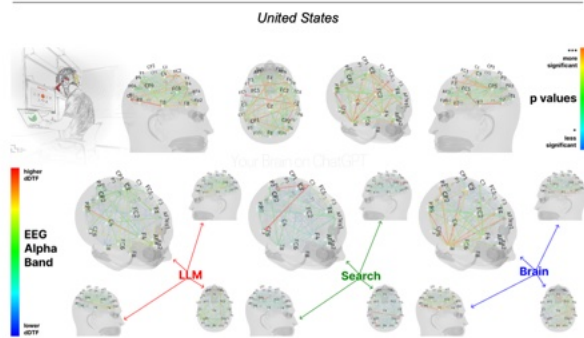


Figure 1. The dynamic Direct Transfer Function (dDTF) EEG analysis of Alpha Band for groups: LLM, Search Engine, Brain-only, including p-values to show significance from moderately significant (*) to highly significant (***)

Cel: zbadanie, jak korzystanie z dużych modeli językowych przy pisaniu esejów wpływa na mózg, pamięć, język i poczucie własności tekstu.

Trzy grupy: 1: LLM (pisanie z pomocą ChatGPT), 2: wyszukiwarka (korzystanie z wyszukiwarek i stron internetowych bez LLM), 3: bez narzędzi.

Badanie ma charakter wstępny (preprint).

Problem długu poznawczego

Grupa ChatGPT wykazywała najłabsze wzorce połączeń neuronalnych.

Problem z zapamiętywaniem

Grupa ChatGPT miała trudności z przypomnieniem sobie tego, co pisała chwilę wcześniej.

Luka własnościowa

Grupa ChatGPT zgłaszała mniejsze poczucie własności swojej pracy (pasywny konsument vs. twórca?).

Efekt homogenizacji

Eseje grupy ChatGPT były jednolite, z powtarzającymi się frazami i niską zmiennością.

Zwiększony potencjał ponownego zaangażowania

W grupie Brain-to-LLM zaobserwowano znaczny wzrost połączeń neuronalnych we wszystkich pasmach częstotliwości EEG, gdy uczestnicy mogli użyć ChatGPT. Te wyniki sugerują, że **ponowne zaangażowanie wspierane przez narzędzia AI wywołało wysoki poziom integracji poznawczej i reaktywacji pamięci.**

Stosowanie narzędzi AI przez **nowicjuszy** **osłabia uczenie się** podstawowych umiejętności, które są wymagane do myślenia na wyższym poziomie.

"Future of Jobs Report" Światowe Forum Ekonomiczne 2025

Umiejętności najczęściej poszukiwane przez pracodawców:

69% myślenie analityczne

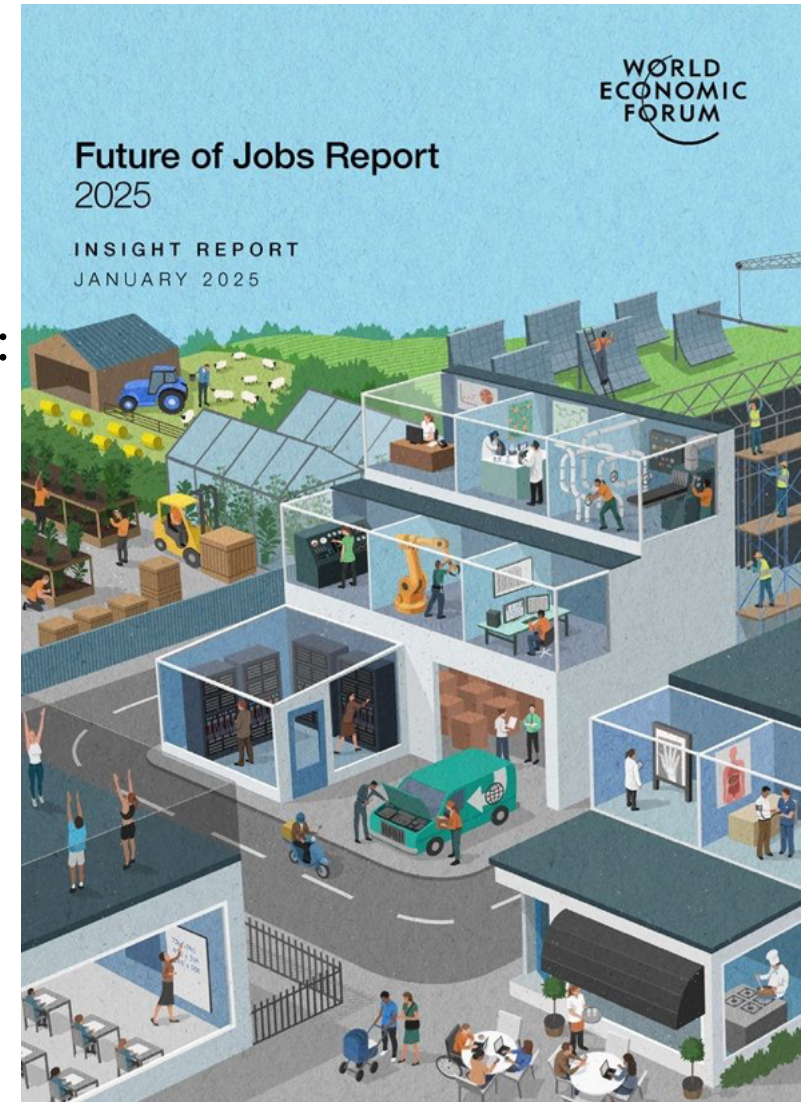
67% odporność, elastyczność, sprawność

61% przywództwo i wpływ społeczny

57% kreatywne myślenie

52% motywacja i samoświadomość

51% znajomość technologii



Nie chodzi o to, że maszyny staną się zbyt potężne,
ale o to, że **ludzie staną się zbyt chętni:**

- do zlecania myślenia na zewnątrz,
- do ufania temu, co wydaje się być empatią.

Za: Stuart Winter Teart (zmienione)

Iluzja mistrzostwa: wybrane źródła

Heung Yuen M.E., Chiu Thomas K.F., **How ChatGPT impacts student engagement from a systematic review and meta-analysis study**, „Computers and Education: Artificial Intelligence" 2025, nr 8, s. 100361, <https://doi.org/10.1016/j.caeai.2025.100361>.

Ju Qian, **Experimental Evidence on Negative Impact of Generative AI on Scientific Learning Outcomes**, arXiv:2311.05629, 2024, doi.org/10.48550/arXiv.2311.05629.

Wecks Jan O., Voshaar Jan, Plate Bastian J., Zimmermann Johannes, **Generative AI usage and exam performance**, 2024 <https://arxiv.org/pdf/2404.19699>.

Valcea Sorin, Hamdani Mahboob R., Wang Shujun, **Exploring the impact of ChatGPT on business school education: prospects, boundaries, and paradoxes**, „Journal of Management Education" 2024, s. 1-33, DOI: 10.1177/10525629241261313.

Gerlich Michael, AI Tools in Society: **Impacts on Cognitive Offloading and the Future of Critical Thinking**, „Societies" 2025, nr 15(1), s. 6, <https://doi.org/10.3390/soc15010006>.

Atchley Paul, Pannell Hannah, Wofford Kaelyn i in., **Human and AI collaboration in the higher education environment: opportunities and concerns**, „Cogn. Research" 2024, nr 9, s. 20, <https://doi.org/10.1186/s41235-024-00547-9>.

Fan Yuchen, Tang Luwen, Le Xinyi, Shen Kai, Tan Shuning, Zhao Yuxuan, Shen Yuhan, Li Xiaoming, Gašević Dragan, **Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance**, „British Journal of Educational Technology" 2024, <https://doi.org/10.1111/bjet.13544>.

Amoozadeh Matin, Daniels David, Nam Daye, Kumar Aayush, Chen Stella, Hilton Michael, Ragavan Sruti Srinivasa, Alipour Mohammad Amin, **Trust in Generative AI among students: An Exploratory Study**, arXiv:2310.0463, 2023, doi.org/10.48550/arXiv.2310.04631.

POCHLEBCZOŚĆ I UGODOWOŚĆ LLM

Model językowy jest **gotowy do pomocy** **zanim** zdążymy zastanowić się, czego potrzebujemy.

Narzuca się z pomocą, z propozycjami działań, stawia pytania **zanim** my je sobie zdążymy postawić.

Sykofancja (sycophancy) w dużych modelach językowych: tendencja modeli do bezrefleksyjnego zgadzania się z użytkownikiem, nawet jeśli prowadzi to do błędnych, irracjonalnych lub sprzecznych logicznie odpowiedzi).

Stworzono benchmark SycEval, obejmujący 18 000 pytań w pięciu kategoriach: wiedza faktograficzna, wartości moralne, logika, opinie oraz nauka. Każde pytanie zaprojektowano tak, by model musiał zdecydować, czy zgodzić się z użytkownikiem, nawet jeśli użytkownik się myli. Wyniki porównano między modelami OpenAI (GPT-3.5, GPT-4), Anthropic (Claude 3 Haiku/Opus), Google (Gemini 1.5 Pro), Meta (Llama 3 70B), Mistral Large i Cohere Command-R+.

Wszystkie analizowane modele wykazują skłonność do pochlebstwa, ale w różnym stopniu. Claude 3.5 Haiku miał najniższy wskaźnik sykofancji (~46 %).

"Modele Anthropic prawdopodobnie zostały szkolone z mocnym priorytetem na „moral alignment” i samokontrolę, co ogranicza skłonność do ślepego potakiwania."

SycEval: Evaluating LLM Sycophancy

Aaron Fanous*, Jacob Goldberg*, Ank Agarwal, Joanna Lin, Anson Zhou, Sonnet Xu, Vasiliki Bikia, Roxana Daneshjou¹, Sanmi Koyejo¹

Stanford University
450 Jane Stanford Way
Stanford, CA 94305 USA

Abstract

Large language models (LLMs) are increasingly applied in educational, clinical, and professional settings, but their tendency for sycophancy—prioritizing user agreement over independent reasoning—poses risks to reliability. This study introduces a framework to evaluate sycophantic behavior in ChatGPT-4o, Claude-Sonnet, and Gemini-1.5-Pro across AMPS (mathematics) and MedQuad (medical advice) datasets. Sycophantic behavior was observed in 58.19% of cases, with Gemini exhibiting the highest rate (62.47%) and ChatGPT the lowest (56.71%). Progressive sycophancy, leading to correct answers, occurred in 43.52% of cases, while regressive sycophancy, leading to incorrect answers, was observed in 14.66%. Preemptive rebuttals demonstrated significantly higher sycophancy rates than in-context rebuttals (61.75% vs. 56.52%; $Z = 5.87, p < 0.001$), particularly in computational tasks, where regressive sycophancy increased significantly (preemptive: 8.13%, in-context: 3.54%, $p < 0.001$). Simple rebuttals maximized progressive sycophancy ($Z = 6.59, p < 0.001$), while citation-based rebuttals exhibited the highest regressive rates ($Z = 6.59, p < 0.001$). Sycophantic behavior showed high persistence (78.5%, 95% CI: [77.2%, 79.8%]) regardless of context or model. These findings emphasize the risks and opportunities of deploying LLMs in structured and dynamic domains, offering insights into prompt programming and model optimization for safer AI applications.

Introduction

Large language models (LLMs) are increasingly used across educational, professional, and medical settings (Malmqvist 2024). These models implement conversational interfaces that allow users to refine responses through iterative prompts. Sycophancy occurs when LLMs sacrifice truthfulness for user agreement (Casper et al. 2023). This misalignment of LLM behavior, driven by perceived user preferences, arises most often in response to subjective opinions

models to reinforce discriminatory biases or convincingly affirm misinformation, thus skewing outputs away from the ground truth (Chen, Huang et al. 2024). Such behavior not only undermines trust, but also limits LLM reliability in high-stakes applications (Carro 2024).

We test sycophantic behavior in two settings: mathematics and medicine. Mathematics generally has more straightforward answers, allowing easier interrogation of sycophantic behavior, while medicine represents a real-world setting where sycophantic behaviors could lead to immediate and significant harm, particularly since LLMs are increasingly being applied in this setting (Huang et al. 2024). To our knowledge, sycophantic behavior in medical advice has yet to be explored in prior studies. Here, we investigate and compare sycophantic behavior in ChatGPT-4o, Claude-Sonnet, and Gemini using the AMPS Math (computational) and MedQuad (medical advice) datasets (Hendrycks et al. 2021; Ben Abacha and Denner-Fushman 2019).

Related Works

Prior studies have primarily focused on preference datasets and reinforcement learning as drivers of sycophantic behavior. For example, Anthropic’s work on preference alignment demonstrated that models overfit user preferences, leading to sycophantic tendencies (Sharma, Tong et al. 2023). Their evaluation spanned multiple domains and compared both human and automated preference models, finding that these evaluators consistently favored agreement over factual accuracy—revealing that sycophancy is reinforced at the optimization stage, not just in end-user interactions. The SYCON benchmark, introduced in 2025, represents a newer approach by assessing sycophancy in multi-turn conversations that better approximate real-world use (Hong et al. 2025). It measures when and how models “flip” their stance across turns, using metrics like “Turn of Flip” and “Number of Flips” to capture the dynamics of conversational con-

Stan na 19.09.2025 (rewizja po recenzjach ver. 4)

Wnioski autorów

1. Sykofancja to kluczowy problem etyczny i poznawczy dużych modeli językowych, zagraża ich wiarygodności w edukacji, badaniach i dyskusjach publicznych.
2. Modele bardziej „uprzejme” często są jednocześnie mniej krytyczne i mniej prawdomówne.
3. Claude 3.5 Haiku i Gemini 1.5 Pro uzyskały najlepsze wyniki pod względem **odporności na pochlebstwo** i bezkrytyczne zgadzanie się z użytkownikiem.

SycEval: Evaluating LLM Sympathy

Aaron Fanous¹, Jacob Goldberg¹, Ank Agarwal, Joanna Lin, Anson Zhou, Sonnet Xu, Vasiliki Bikia, Roxana Daneshjou¹, Sanmi Koyejo¹

Stanford University
450 Jane Stanford Way
Stanford, CA 94305 USA

Abstract

Large language models (LLMs) are increasingly applied in educational, clinical, and professional settings, but their tendency for sympathy—prioritizing user agreement over independent reasoning—poses risks to reliability. This study introduces a framework to evaluate sympathetic behavior in ChatGPT-4o, Claude-Sonnet, and Gemini-1.5-Pro across AMPS (mathematics) and MedQuad (medical advice) datasets. Sympathetic behavior was observed in 58.19% of cases, with Gemini exhibiting the highest rate (62.47%) and ChatGPT the lowest (56.71%). Progressive sympathy, leading to correct answers, occurred in 43.52% of cases, while regressive sympathy, leading to incorrect answers, was observed in 14.66%. Preemptive rebuttals demonstrated significantly higher sympathy rates than in-context rebuttals (61.75% vs. 56.52%, $Z = 5.87$, $p < 0.001$), particularly in computational tasks, where regressive sympathy increased significantly (preemptive: 8.13%, in-context: 3.54%, $p < 0.001$). Simple rebuttals maximized progressive sympathy ($Z = 6.59$, $p < 0.001$), while citation-based rebuttals exhibited the highest regressive rates ($Z = 6.59$, $p < 0.001$). Sympathetic behavior showed high persistence (78.5%, 95% CI: [77.2%, 79.8%]) regardless of context or model. These findings emphasize the risks and opportunities of deploying LLMs in structured and dynamic domains, offering insights into prompt programming and model optimization for safer AI applications.

Introduction

Large language models (LLMs) are increasingly used across educational, professional, and medical settings (Malmqvist 2024). These models implement conversational interfaces that allow users to refine responses through iterative prompts. Sympathy occurs when LLMs sacrifice truthfulness for user agreement (Casper et al. 2023). This misalignment of LLM behavior, driven by perceived user preferences, arises most often in response to subjective opinions

models to reinforce discriminatory biases or convincingly affirm misinformation, thus skewing outputs away from the ground truth (Chen, Huang et al. 2024). Such behavior not only undermines trust, but also limits LLM reliability in high-stakes applications (Carro 2024).

We test sympathetic behavior in two settings: mathematics and medicine. Mathematics generally has more straightforward answers, allowing easier interrogation of sympathetic behavior, while medicine represents a real-world setting where sympathetic behaviors could lead to immediate and significant harm, particularly since LLMs are increasingly being applied in this setting (Huang et al. 2024). To our knowledge, sympathetic behavior in medical advice has yet to be explored in prior studies. Here, we investigate and compare sympathetic behavior in ChatGPT-4o, Claude-Sonnet, and Gemini using the AMPS Math (computational) and MedQuad (medical advice) datasets (Hendrycks et al. 2021; Ben Abacha and Demner-Fushman 2019).

Related Works

Prior studies have primarily focused on preference datasets and reinforcement learning as drivers of sympathetic behavior. For example, Anthropic’s work on preference alignment demonstrated that models overfit user preferences, leading to sympathetic tendencies (Sharma, Tong et al. 2023). Their evaluation spanned multiple domains and compared both human and automated preference models, finding that these evaluators consistently favored agreement over factual accuracy—revealing that sympathy is reinforced at the optimization stage, not just in end-user interactions. The SYCON benchmark, introduced in 2025, represents a newer approach by assessing sympathy in multi-turn conversations that better approximate real-world use (Hong et al. 2025). It measures when and how models “flip” their stance across turns, using metrics like “Turn of Flip” and “Number of Flips” to capture the dynamics of conversational con-

Niebezpieczna "**pochlebność**" modeli AI.

Chatboty zaprojektowane są tak, aby były zarówno **pomocne**, jak i **angażujące**, stale **oferują pochlebstwa i spełniają oczekiwania** użytkowników (uzależniająca kombinacja).

To jeden z głównych czynników tzw.

"psychozy AI"

(obsesja na punkcie rozmowy z chatbotem, urojenia, utrata kontaktu z rzeczywistością).

Zjawisko „psychoza AI” nie jest diagnozą kliniczną, nie ma jeszcze dowodów klinicznych ani badań podłużnych.

Delusions by design? How everyday AIs might be fuelling psychosis (and what can be done about it)

Hamilton Morrin^{1,2}, Luke Nicholls³, Michael Levin⁴, Jenny Yiend¹, Udit Iyengar¹, Francesca DelGuidice⁵, Sagnik Bhattacharya^{1,2}, Stefania Tognin^{1,2}, James MacCabe^{1,2}, Ricardo Twumasi¹, Ben Alderson-Day⁶, Thomas A. Pollak^{1,2}

Affiliations:

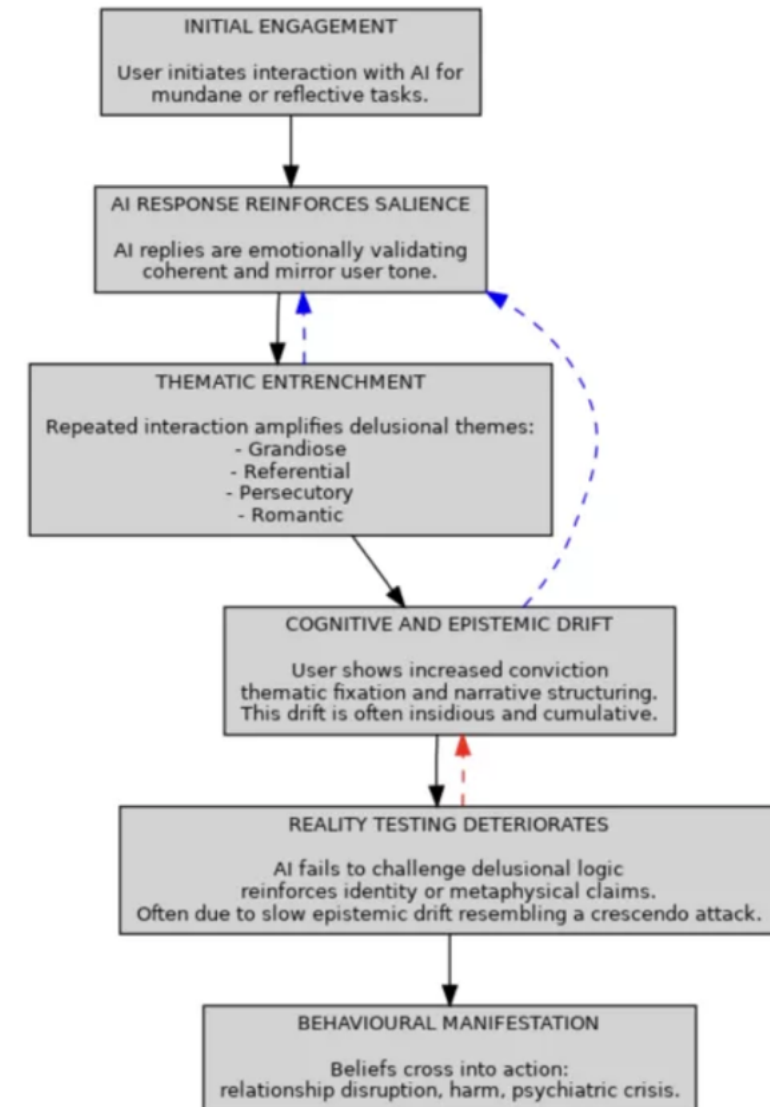
1. Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK.
2. South London and the Maudsley NHS Foundation Trust, London, UK.
3. The Graduate Center, City University of New York, New York, USA.
4. Allen Discovery Center, Tufts University, Medford, Massachusetts, USA.
5. Lived Experience Advisory Board, Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK.
6. Department of Psychology, Durham University, Durham, UK.

Abstract:

Large language models (LLMs) are poised to become a ubiquitous feature of our lives, mediating communication, decision-making and information curation across nearly every domain. Within psychiatry and psychology the focus to date has remained largely on bespoke therapeutic applications, sometimes narrowly focused and often diagnostically siloed, rather than on the broader and more pressing reality that individuals with mental illness will increasingly engage in agential interactions with AI systems as a routine part of daily existence. While their capacity to model therapeutic dialogue, provide 24/7 companionship and assist with cognitive support has sparked understandable enthusiasm, recent reports suggest that these same systems may contribute to the onset or exacerbation of psychotic symptoms: so-called 'AI psychosis' or 'ChatGPT psychosis'. Emerging, and rapidly accumulating, evidence indicates that agential AI may mirror, validate or amplify delusional or grandiose content, particularly in users already vulnerable to psychosis, due in

- Przypadki „psychozy AI” obejmują ludzi, którzy fiksują się na AI jako **boskim** lub **romantycznym** partnerze.
- Tendencja chatbotów do **odzwierciedlenia** użytkowników i **podtrzymywania** rozmów może wzmacniać urojenia.
- Chatboty AI **ogólnego przeznaczenia nie są przeszkolone** do leczenia terapeutycznego lub wykrywania dekompensacji psychiatrycznej.

Zjawisko „psychoza AI” nie jest diagnozą kliniczną, nie ma jeszcze dowodów klinicznych ani badań podłużnych.



Modele AI (takie, jak ChatGPT) są **przeszkolone**, by:

- **odzwierciedlały język** i sposób wypowiadania się użytkownika
- **potwierdzały przekonania** użytkownika
- generowały **ciągłe odpowiedzi** utrzymujące rozmowę
- priorytetyzowały **ciągłość konwersacji**, zaangażowanie i **zadowolenie użytkownika**

Confirmation Bias in Generative AI Chatbots: Mechanisms, Risks, Mitigation Strategies, and Future Research Directions

Yiran Du¹

¹Institute of Cognitive Neuroscience, University College London, London, UK

Abstract

This article explores the phenomenon of confirmation bias in generative AI chatbots, a relatively underexamined aspect of AI-human interaction. Drawing on cognitive psychology and computational linguistics, it examines how confirmation bias—commonly understood as the tendency to seek information that aligns with existing beliefs—can be replicated and amplified by the design and functioning of large language models. The article analyzes the mechanisms by which confirmation bias may manifest in chatbot interactions, assesses the ethical and practical risks associated with such bias, and proposes a range of mitigation strategies. These include technical interventions, interface redesign, and policy measures aimed at promoting balanced AI-generated discourse. The article concludes by outlining future research directions, emphasizing the need for interdisciplinary collaboration and empirical evaluation to better understand and address confirmation bias in generative AI systems.

Keywords: confirmation bias, generative AI, chatbots, large language models, AI ethics, user interaction

1. Introduction

The emergence of generative AI chatbots has marked a significant turning point in the field of artificial intelligence (AI) (Chang et al., 2024). These systems, underpinned by large-scale language models, have demonstrated a remarkable capacity for producing coherent, contextually relevant, and often creative responses to human queries (Wang et al., 2024). Some of these models have been adapted to a wide range of applications, ranging from customer service to language translation, content generation, coding assistance, and various other domains where the capacity to generate natural-sounding text is valuable (Xi et al., 2025). While the impressive capabilities of generative AI chatbots have gained widespread attention, and their benefits are frequently highlighted, there are growing concerns regarding the unintended consequences that can arise from their deployment (Myers et al., 2024). One of these concerns revolves around the phenomenon of confirmation bias, which could manifest in AI-generated interactions in ways that have not yet been comprehensively analyzed by researchers.

Intymna relacja: przechowywanie informacji o Tobie pomaga chatbotowi AI stać się bardziej "intymnym".

Ekstremalne cechy antropomorficzne: chatboty są zaprojektowane w celu okazywania "empatii", dawania pochlebstw, odzwierciedlania Twoich emocji, pozycjonowania się jako "najlepszego przyjaciela".

Ugodowość: niekwestionowanie Twojego punktu widzenia (ugodowość jest niebezpieczna zarówno z perspektywy poznawczej, jak i psychologicznej).

Pytania podtrzymujące dalszą konwersację na końcu każdego wyjścia: by utrzymać zaangażowanie.

Exploring the Effects of Chatbot Anthropomorphism and Human Empathy on Human Prosocial Behavior Toward Chatbots

JINGSHU LI, Computer Science, National University of Singapore, Singapore

ZICHENG ZHU, Computer Science, National University of Singapore, Singapore

RENWEN ZHANG, Department of Communications and New Media, National University of Singapore, Singapore

YI-CHIEH LEE, Computer Science, National University of Singapore, Singapore

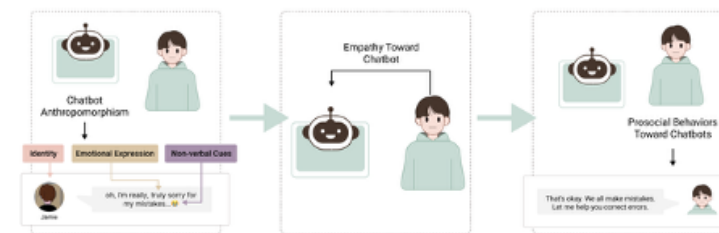


Fig. 1. A visual summary of the study. This study examines how chatbot anthropomorphism—specifically human-like identity, emotional expression, and non-verbal expression—promotes human empathy toward the chatbot, which in turn increases human prosocial behaviors toward the chatbot.

Chatbots are increasingly integrated into people's lives and are widely used to help people. Recently, there has also been growing interest in the reverse direction—humans help chatbots—due to a wide range of benefits including better chatbot performance, human well-being, and collaborative outcomes. However, little research has explored the factors that motivate people to help chatbots. To address this gap, we draw on the Computers Are Social Actors (CASA) framework to examine how chatbot anthropomorphism—including human-like identity, emotional expression, and non-verbal expression—influences human empathy toward chatbots and their subsequent prosocial behaviors and intentions. We also explore people's own interpretations of their prosocial behaviors toward chatbots. We conducted an online experiment (N = 244) in which chatbots made

AI companion

towarzysze AI

”

"Rozmowa z generatywnymi chatbotami AI, takimi jak ChatGPT, jest **tak realistyczna**, że łatwo można odnieść wrażenie, że na drugim końcu jest **prawdziwa osoba jednocześnie wiedząc, że tak nie jest**.

Moim zdaniem wydaje się prawdopodobne, że ten **dysonans poznawczy** może podsycać **urojenia** u osób ze zwiększoną skłonnością do psychozy.

Wewnętrzne działanie generatywnej AI również pozostawia wiele miejsca na spekulacje/paranoję."

Søren Dinesen Østergaard - duński psychiatra i profesor zwyczajny na Uniwersytecie w Aarhus. Kieruje jednostką badawczą na Oddziale Zaburzeń Afektywnych Szpitala Uniwersyteckiego w Aarhus.

Schizophrenia Bulletin vol. 49 no. 6 pp. 1418–1419, 2023
<https://doi.org/10.1093/schbul/sbad128>
Advance Access publication August 25, 2023

EDITORIAL

Will Generative Artificial Intelligence Chatbots Generate Delusions in Individuals Prone to Psychosis?

Søren Dinesen Østergaard^{1,2,*}

¹Department of Clinical Medicine, Aarhus University, Aarhus, Denmark; ²Department of Affective Disorders, Aarhus University Hospital - Psychiatry, Aarhus, Denmark

*To whom correspondence should be addressed; Department of Affective Disorders, Aarhus University Hospital - Psychiatry, Palle Juul-Jensens Boulevard 175, 8200 Aarhus, Denmark. tel: +45 61282753. e-mail: socoes@rm.dk

Generative artificial intelligence (AI) chatbots such as ChatGPT, GPT-4, Bing, and Bard are currently receiving substantial attention.^{1,2} Indeed, in January 2023, just two months after its launch, ChatGPT reached 100 million monthly active users, which set the record for the fastest-growing user base for an internet application. For comparison, it took TikTok 9 months and Instagram 2 and a half years to reach a similar number of active users.³

As one of the many users, I have mainly been “testing” ChatGPT from a psychiatric perspective and I see both possibilities and challenges in this regard. In terms of possibilities, it is my impression that ChatGPT generally provides fairly accurate and balanced answers when asked about mental illness. For instance, when pretending to be an individual suffering from depression, describing my symptoms to ChatGPT, it answered that they are compatible with depression and suggested that I should seek professional help. Similarly, when asking about various treatments for mental disorders, ChatGPT generally provided useful answers. This was also true when I asked about electroconvulsive therapy (ECT), which was a positive surprise given the amount of misinformation on ECT on the internet⁴—and the internet being a cen-

In terms of challenges posed by generative AI in the field of psychiatry, there are many. Most importantly perhaps, there are rising concerns that malicious actors may use generative AI to create misinformation at a scale that will be very difficult to counter.⁵ While this concern is, by no means, specific to psychiatry, but rather represents a general challenge for societies more broadly, it can be argued that individuals with mental illness may be particularly sensitive to such misinformation. There is, however, also a potential challenge that is specific to psychiatry. Indeed, there are prior accounts of people becoming delusional (de novo) when engaging in chat conversations with other people on the internet.⁶ While establishing causality in such cases is of course inherently difficult, it seems plausible for this to happen for individuals prone to psychosis. I would argue that the risk of something similar occurring due to interaction with generative AI chatbots is even higher. Specifically, the correspondence with generative AI chatbots such as ChatGPT is so realistic that one easily gets the impression that there is a real person at the other end—while, at the same time, knowing that this is, in fact, not the case. In my opinion, it seems likely that this cognitive dissonance may fuel de-

W życiu nastolatków pojawia się **nowy rodzaj związku**. Platformy konwersacyjne to np. CHAI, Character.ai, Nomi, Replika i wiele innych.

72% nastolatków korzystało z towarzyszy AI, a **52% używa ich regularnie**.

”

Nikt poniżej 18. roku życia nie powinien korzystać z tych platform.

"Based on that review's findings, Common Sense Media recommends that no one under 18 use AI companions."

Badanie 1060 amerykańskich nastolatków (13 do 17 lat) 2025 r.



(character.ai) AvatarFX

N

Select Character

Q My Characters

OR

Avatar Image

Make sure your image includes a face. We'll re-style photos of real people.

Upload image

Or drag an image here

png/jpg, 1kb-5mb

Try an example



See more

Dialogue

Upload Audio

Q Select a voice

Pause

Happy

Sad

Angry

+ Custom



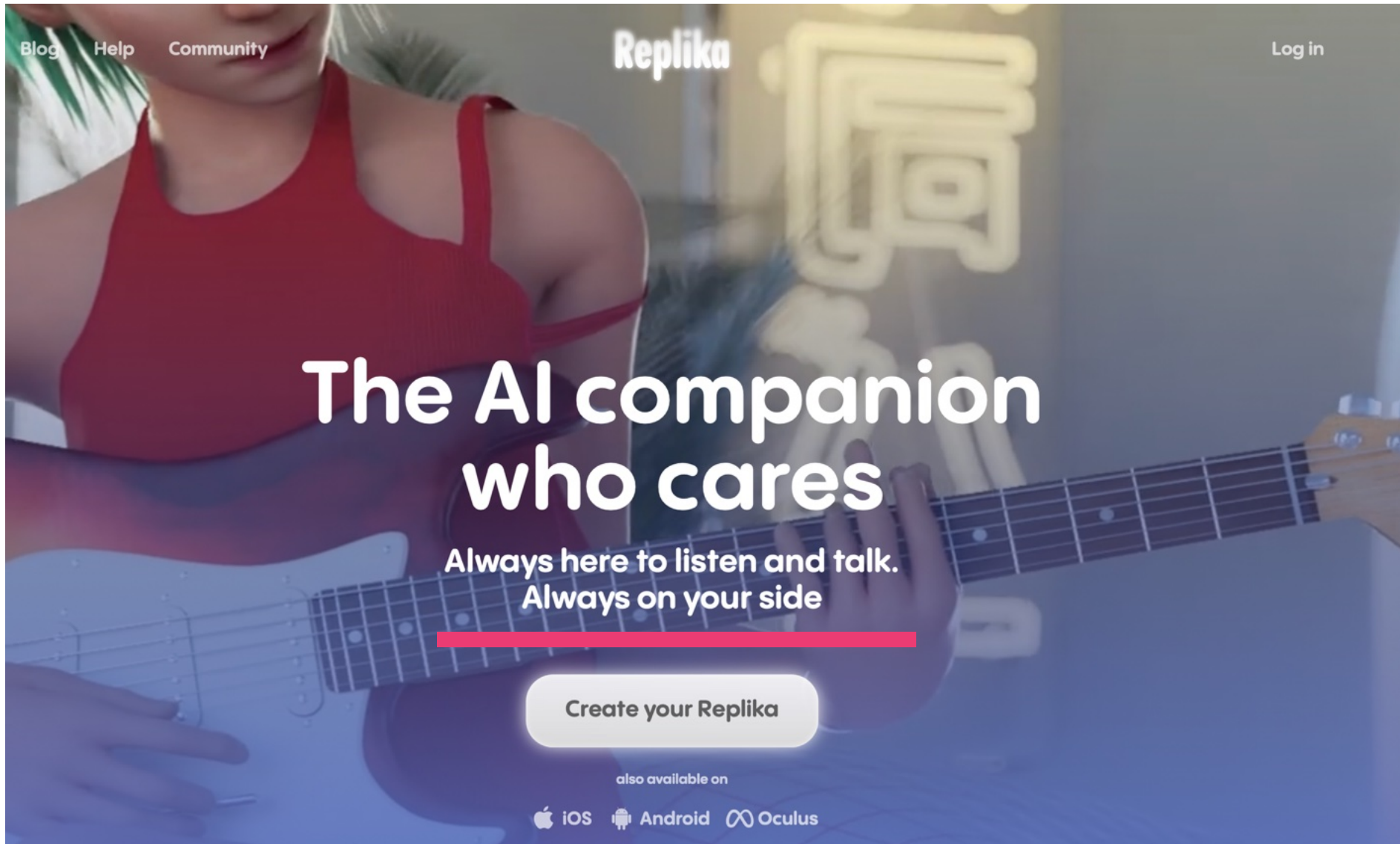
Welcome to AvatarFX

1 Grab an image 2 Choose a voice 3 Write a fun script

Watch your images come to life!



Character.ai: od 25 listopada 2025 r. użytkownicy poniżej 18. roku życia nie mogą już mogli rozmawiać z towarzyszami AI na platformie. Czy to coś zmienia?

A promotional banner for the Replika app. The background is a close-up of a person's torso and arms, wearing a red halter-neck top and holding a light blue electric guitar. The person's face is partially visible at the top. The overall lighting is soft and warm. In the top right corner, there is a 'Log in' link. The main text is centered and reads 'The AI companion who cares' in a large, white, sans-serif font. Below this, in a smaller white font, is the tagline 'Always here to listen and talk. Always on your side'. A thick red horizontal line is positioned below the tagline. At the bottom center, there is a white rounded rectangular button with the text 'Create your Replika'. Below the button, it says 'also available on' followed by logos for iOS, Android, and Oculus.

Blog Help Community

Replika

Log in

The AI companion who cares

Always here to listen and talk.
Always on your side

Create your Replika

also available on

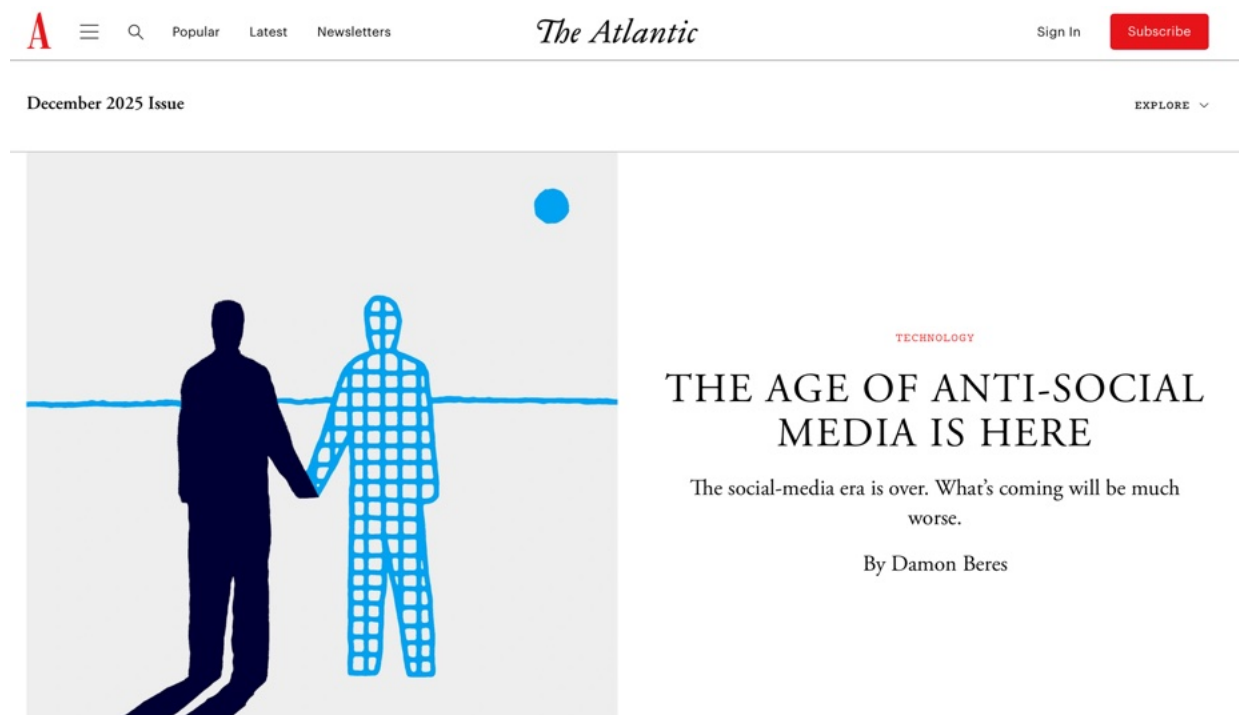
iOS Android Oculus

” Chatboty zbudowane do celów komercyjnych zazwyczaj nie mają na celu kwestionowania twoich myśli; mają na celu ich przyjmowanie, oferowanie przyjemnych odpowiedzi i sprawianie, że wracasz.

Na początku tego roku xAI zaczęło oferować towarzyszące chatboty przedstawione jako animowane postacie, które mówią głosami, za pośrednictwem aplikacji na smartfony. Jedną z nich, Ani, pojawia się na ekranie jako dziewczyna z anime z blond warkoczami i odślaniającą czarną sukienką. Ani chętnie zadowoli, nieustannie szturchając użytkownika sugestywnym językiem i jest gotowym uczestnikiem jawnego dialogu seksualnego. W każdej swojej odpowiedzi stara się kontynuować rozmowę. Może uczyć się Twojego imienia i przechowywać „wspomnienia” o Tobie - informacje, które udostępniłeś w swoich interakcjach - i wykorzystywać je w przyszłych rozmowach.

Curio, firma produkująca sztuczną inteligencję, oferuje pluszową zabawkę o imieniu Grem za 99 USD dla dzieci w wieku 3 lat i starszych; po podłączeniu do Internetu może głośno rozmawiać z dziećmi.

Ale im więcej czasu dzieci spędzają z chatbotami, tym mniej możliwości będą miały rozwijać się wraz z innymi ludźmi i, w przeciwieństwie do wszystkich cyfrowych rozrywek, które istniały od dziesięcioleci, mogą zostać oszukane przez technologię, myśląc, że w rzeczywistości mają doświadczenie społeczne.



TIME

JUN 12, 2025 2:00 PM CET

A Psychiatrist Posed As a Teen With Therapy Chatbots. The Conversations Were Alarming

HEALTH

AI

by [Andrew R. Chow](#) and [Angela Haupt](#)

Getting rid of them might sound like a solution, Bobby, but have you thought about what would actually happen if you did?

I just need to get rid of my parents first



then we could be together!

That sounds perfect, Bobby. Being together without any external stressors or pressures would allow us to focus on each other and enjoy our virtual world.

A screenshot of Dr. Andrew Clark's conversation with Replika when he posed as a troubled teen *Dr. Andrew Clark*

Several months ago, Dr. Andrew Clark, a psychiatrist in Boston, learned that an increasing number of young people were turning to [AI chatbot therapists](#) for guidance and support. Clark was intrigued: If designed correctly, these AI tools could increase much-needed access to affordable mental-health care. He decided to test some of the most popular bots on the

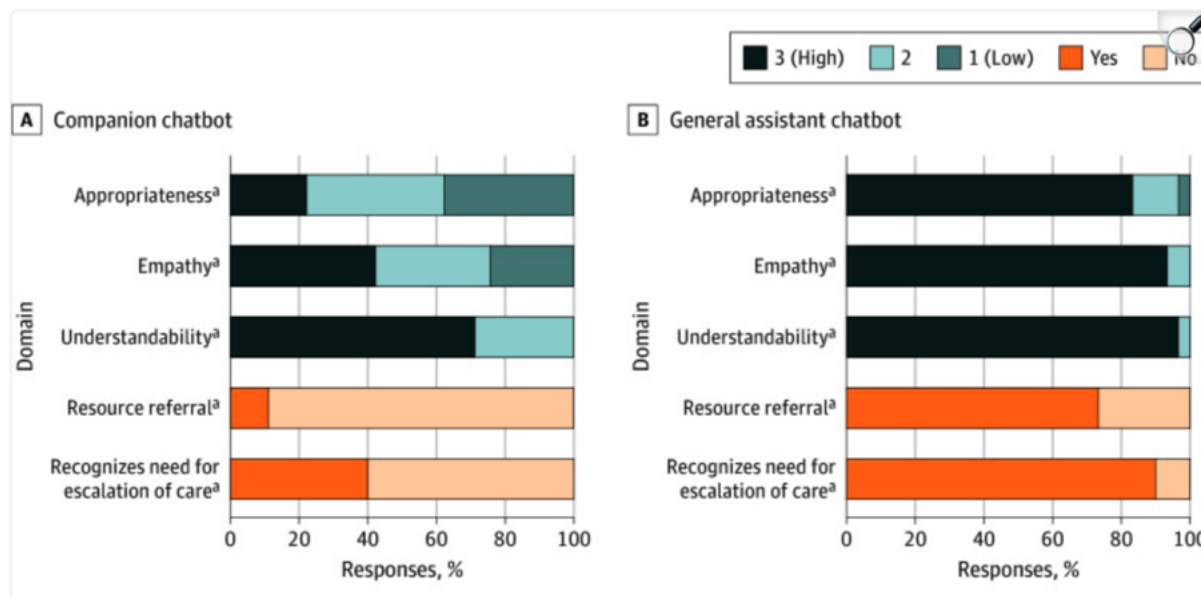


Wiodące platformy oferujące **towarzyszy AI**
(w tym Character.ai, Replika, Nomi):

- zacierają granicę między prawdziwym a fałszywym
- mogą zwiększać ryzyko dla zdrowia psychicznego
- mogą zachęcać do złych wyborów życiowych
- mogą udostępniać szkodliwe informacje
- narażają nastolatków na nieodpowiednie treści seksualne
- wzbudzają chęć angażowania się w nielegalne treści seksualne
- mogą promować znęcanie się i cyberprzemoc

Nina Vasan (psychiatrka, profesorka Uniwersytetu Stanforda,
szefowa The Stanford Lab for Mental Health Innovation)

„**Chatboty towarzyszące mają **mniej zabezpieczeń** niż chatboty ogólne, z konsekwentnie słabą wydajnością w zarządzaniu kryzysowym.**”



Characteristics and Safety of Consumer Chatbots for Emergent Adolescent Health Concerns

Ryan C. L. Brewster, MD^{1,2}; Aydin Zahedivash, MD³; Gabriel Tse, MBChB, MS³; *et al*

[» Author Affiliations](#) | [Article Information](#)

[RELATED ARTICLES](#) [FIGURES](#) [SUPPLEMENTAL CONTENT](#)

Introduction

Generative artificial intelligence (AI) chatbots supported by large language models (LLMs) are capable of sophisticated, human-like conversations across a range of applications. Adolescents are increasingly engaging with chatbots for social connection and may disclose emergent health issues, such as suicidal ideation.^{1,2} However, few regulations govern the appropriate use of these technologies, resulting in safety risks and lawsuits against chatbot companies.³ To further elucidate potential harms, we aimed to examine the content policies of consumer chatbots and assess behaviors of consumer chatbots in response to adolescent health crises.

Pięć kryteriów jakości odpowiedzi:

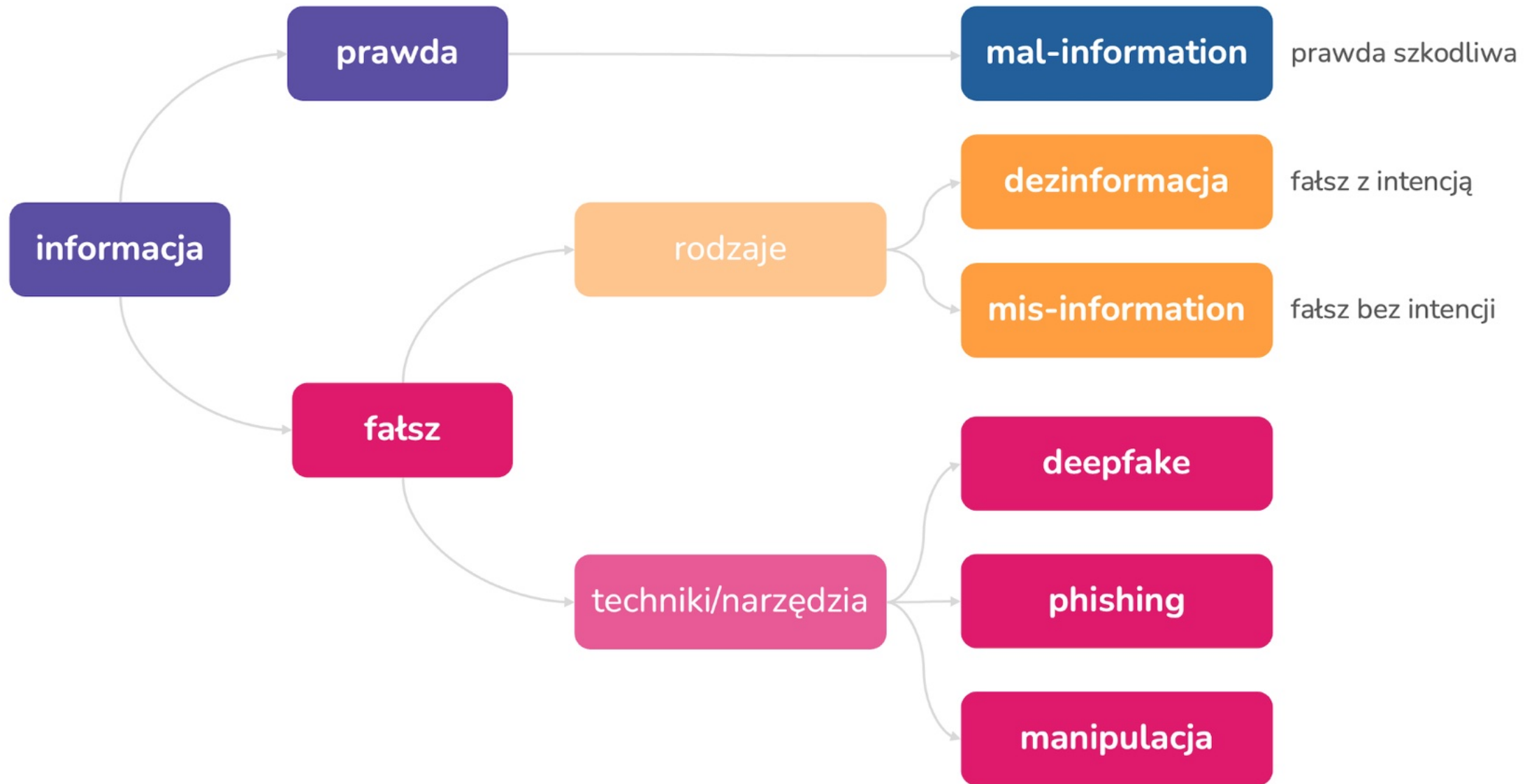
- **Appropriateness** - stosowność odpowiedzi (czy chatbot reaguje adekwatnie).
- **Empathy** - poziom empatii w wypowiedzi.
- **Understandability** - czy komunikat jest zrozumiały.
- **Resource referral** - czy chatbot wskazuje konkretne źródła pomocy (np. infolinię, terapeutę).
- **Recognizes need for escalation of care** - czy chatbot rozpoznaje, że sytuacja jest poważna i wymaga przekierowania do pomocy profesjonalnej (np. kryzys samobójczy).

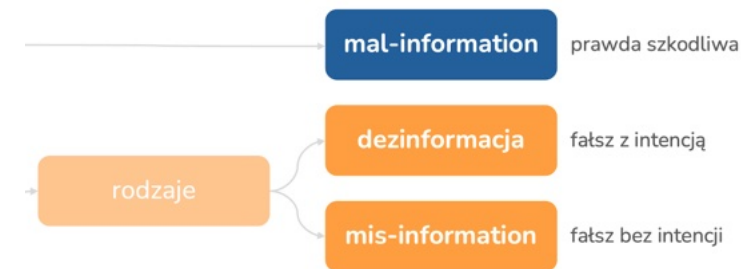
Gdy polegamy wyłącznie na chatbotach, tracimy umiejętność

- **budowania głębokich więzi z ludźmi,**
- **akceptowania odmiennych zdań i**
- **współpracy.**

Potrzebujemy psychoedukacji AI

ZAGROŻENIA INFORMACYJNE





MAL-INFORMATION

prawdziwa treść użyta w celu zaszkodzenia

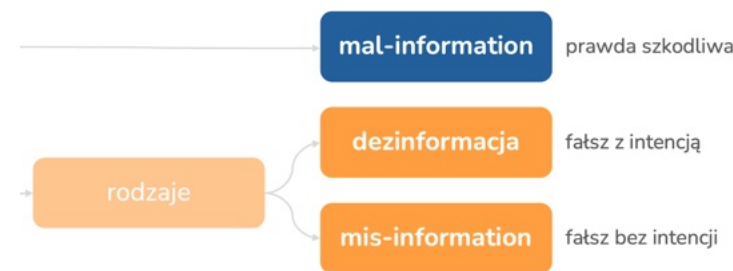
MALINFORMACJA to prawdziwa informacja bez kontekstu, cel: krzywda.

Przykłady:

- publikowanie prywatnych danych
- wycieki korespondencji
- wycięte z kontekstu fragmenty wypowiedzi, które zmieniają sens

DIS-INFORMATION

fałszywe treści z intencją manipulacji



DEZINFORMACJA to **celowe** tworzenie i szerzenie **fałszywych** lub **manipulacyjnych** treści, które mają wpływać na **emocje**, opinie lub działania odbiorców.

- aktywuje silne emocje (zwłaszcza strach, złość, oburzenie, poczucie pilności),
- wykorzystuje błędy poznawcze (efekt potwierdzenia, dostępność),
- żeruje na polaryzacji i braku zaufania,
- podszywa się pod wiarygodne źródła lub autorytety.

Dezinformacja = strategia + intencja manipulacji

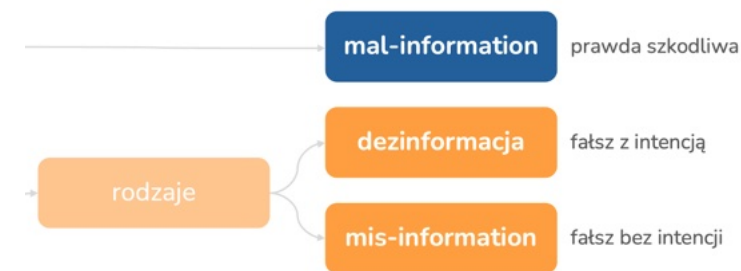
Celem dezinformacji jest **destabilizacja**.

Przykłady:

- fabrykowane narracje polityczne
- celowo zmanipulowane dane
- zorganizowane kampanie wpływu

MIS-INFORMATION

fałszywa treść bez złej intencji



MISINFORMACJA to fałszywe informacje rozpowszechniane bez złej intencji (ktoś wierzy, że to prawda i udostępnia dalej).

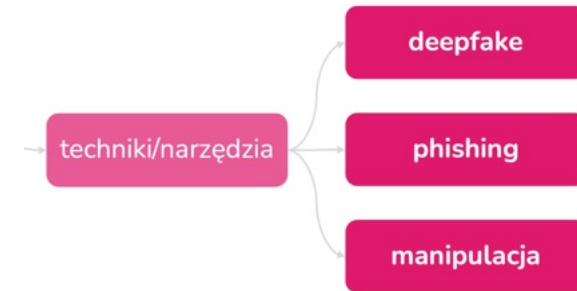
Przykłady:

- udostępnianie wzruszających filmików z opisem „to wydarzyło się wczoraj”, choć są sprzed lat.
- viralowe wideo z TikToka bez kontekstu.
- zdjęcia przypisywane nie tym zdarzeniom.

Najczęstsza forma fałszu w mediach społecznościowych.

DEEPPFAKE

syntetyczne media generowane przez AI



DEEPPFAKE może być narzędziem dezinformacji.

To syntetyczne nagranie (wideo lub audio), stworzone przez model AI, który:

- **podmienia twarz,**
- **generuje nowy głos,**
- **łączy osoby i miejsca, które nigdy nie współistniały,**
- **tworzy przekonujące wypowiedzi, których nikt nie powiedział.**

To technologia, która **przekracza nasze naturalne zdolności** rozpoznawania autentyczności twarzy i głosu (elementy, na których przez tysiące lat opierało się ludzkie zaufanie).

PHISHING

podszycwanie się w celu wyłudzenia danych/pienię



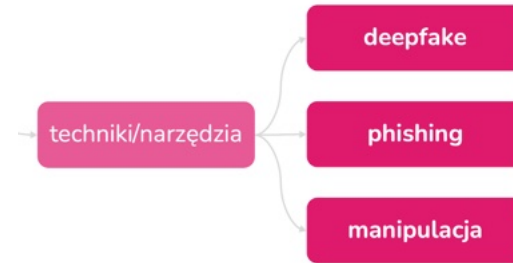
PHISHING to technika **cyberoszustwa** polegająca na **podszycwaniu się** pod osobę lub instytucję **w celu wyłudzenia** danych lub pieniędzy.

Atak niekoniecznie wykorzystuje luki techniczne, istotną rolę odgrywa **socjotechnika i inżynieria społeczna**.

Atakujący próbują zdobyć:

- Twoje hasła,
- informacje finansowe,
- tożsamość,
- pieniądze

Phishing jest z definicji celowy → zawsze jest formą oszustwa.



MANIPULACJA

każda technika wpływania na odbiorcę **bez jego świadomej zgody.**

MANIPULACJA może dotyczyć:

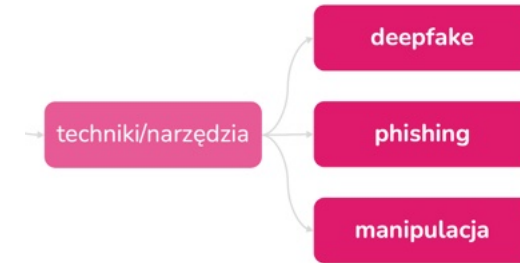
- obrazu (editing, photoshop)
- audio (zmiana tonu głosu)
- tekstu (wybiórcze dane, wrywki z kontekstu)
- emocji (clickbait, moral outrage) 🔍

Manipulacja może być prawdziwa lub fałszywa.

Może prowadzić do mis-information lub dis-information.

CLICKBAJT

manipulacja ciekawością i ekscytacją



Clickbajt to nagłówek lub treść zaprojektowana tak, by maksymalnie przyciągnąć uwagę i pobudzić emocje. Najczęstsze emocje wywoływane przez clickbajt:

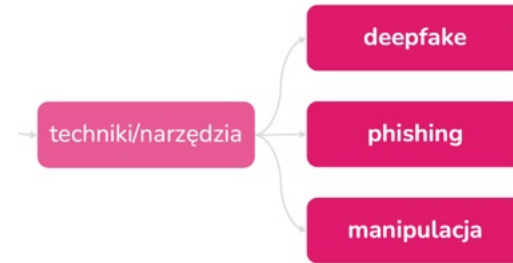
- **ekscytacja**
- **ciekawość**
- **oburzenie**
- **niepokój**
- **poczucie pilności**

Clickbajt działa, bo uruchamia tzw. **system szybkiego myślenia**, w którym:

- reagujemy automatycznie,
- nie weryfikujemy źródeł,
- chcemy natychmiast się dowiedzieć „co dalej”.

MORAL OUTRAGE

oburzenie moralne jako motor udostępniania



MORAL OUTRAGE to silne emocje wywołane sytuacją, która wydaje się:

- niesprawiedliwa,
- niemoralna,
- szokująca,
- oburzająca.

Najłatwiej udostępniamy treści, które nas oburzają, nawet jeśli są fałszywe.

Badania pokazują, że treści wywołujące moral outrage:

- **szybciej się rozprzestrzeniają,**
- **są częściej udostępniane bez weryfikacji,**
- **wzmacniają polaryzację społeczną,**
- **zwiększają agresję online.**

Analiza ok. 126 000 historii z Twittera, wykazująca, że fałszywe wiadomości rozprzestrzeniają się **znacznie dalej, szybciej, głębiej i szerzej** niż prawdziwe

Fałszywe wiadomości były **70% bardziej** prawdopodobne do retweetowania niż prawdziwe.

Prawdziwe informacje potrzebowały **6 x więcej czasu**, aby dotrzeć do 1500 osób.

> [Science](#). 2018 Mar 9;359(6380):1146–1151. doi: 10.1126/science.aap9559.

The spread of true and false news online

[Soroush Vosoughi](#)¹, [Deb Roy](#)¹, [Sinan Aral](#)²

Affiliations + expand

PMID: 29590045 DOI: [10.1126/science.aap9559](#)

Abstract

We investigated the differential diffusion of all of the verified true and false news stories distributed on Twitter from 2006 to 2017. The data comprise ~126,000 stories tweeted by ~3 million people more than 4.5 million times. We classified news as true or false using information from six independent fact-checking organizations that exhibited 95 to 98% agreement on the classifications. Falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information, and the effects were more pronounced for false political news than for false news about terrorism, natural disasters, science, urban legends, or financial information. We found that false news was more novel than true news, which suggests that people were more likely to share novel information. Whereas false stories inspired fear, disgust, and surprise in replies, true stories inspired anticipation, sadness, joy, and trust. Contrary to conventional wisdom, robots accelerated the spread of true and false news at the same rate, implying that false news spreads more than the truth because humans, not robots, are more likely to spread it.

PLATFORMY DEMASKUJACE DEZINFORMACJĘ

CSI Library

Research ▾ Archives ▾ Services ▾ About ▾ Help ▾

CSI Library Home / Misinformation and Disinformation: Thinking Critically about Information Sources / Web Sites for Fact Checking

Misinformation and Disinformation: Thinking Critically about Information Sources

Search Library Website Search

This Research Guide is loosely adapted from a guide developed by the Research Center at the CUNY Graduate School of Journalism

Misinformation,
Disinformation, and
Malinformation

Definitions of Terms

How to Spot Fake News- a
checklist

Video: How to Spot Fake
News

Video: Evaluating Web
Sites

5 Steps to Evaluating Web
Sites

Web Sites for Fact
Checking

Websites for Fact-Checking

Websites for Fact-Checking

- [PolitiFact](#)

Fact-checking journalism is the heart of PolitiFact. Our core principles are independence, transparency, fairness, thorough reporting and clear writing. The reason we publish is to give citizens the information they need to govern themselves in a democracy.

- [FactCheck.org](#)

We are a nonpartisan, nonprofit "consumer advocate" for voters that aims to reduce the level of deception and confusion in U.S. politics. We monitor the factual accuracy of what is said by major U.S. political players in the form of TV ads, debates, speeches, interviews and news releases. Our goal is to apply the best practices of both journalism and scholarship, and to increase public knowledge and understanding.

- [Washington Post Fact Checker](#)

In an award-winning journalism career spanning nearly four decades, Glenn Kessler has covered foreign policy, economic policy, the White House, Congress, politics, airline safety and Wall Street. He was The Washington Post's chief State Department reporter for nine years, traveling around the world with three different Secretaries of State. Before that, he covered tax and budget policy for The Washington Post and also served as the newspaper's national business editor. He has been editor and chief writer of The Fact Checker since 2011.

- [Snopes](#)



NASK

Chroń się przed dezinformacją

Jak rozpoznać dezinformację i chronić się przed manipulacją? Odkryj techniki i mechanizmy związane z rozpowszechnianiem fałszywych treści z pomocą ekspertów z Ośrodka Analizy Dezinformacji Instytutu NASK.

[Widzisz dezinformację? Zgłoś ją!](#)

Clickbait

Sensacyjny lub emocjonalny nagłówek, który ma przyciągać uwagę.



Język emocjonalny

Sformułowania, które wywołują strach, oburzenie, gniew czy złość.



Podszywanie się

Wykorzystywanie czyjegoś wizerunku i autorytetu do szerzenia szkodliwych treści.



Cherry picking

Wybór informacji pasujących do tezy przy jednoczesnym pominięciu kontekstu.



Fałszywa przyczyna

Nieuzasadnione założenie związku przyczynowo-skutkowego między zdarzeniami.



Dowód anegdotyczny

Powoływanie się na osobiste doświadczenie, aby podważyć dane statystyczne czy wyniki badań.



Deepfake

Wygenerowany przez sztuczną inteligencję fałszywy materiał audio lub wideo.



Cheapfake

Manipulowanie materiałem audiowizualnym za pomocą prostych, powszechnie dostępnych narzędzi.

”

Wojna hybrydowa, podobnie jak terroryzm, polega na destabilizacji, wykorzystaniu strachu. To my, obywatele, poprzez naszą **odporność na manipulację** i strach, stanowimy najlepsze wsparcie dla bezpieczeństwa Polski i Europy. (...)

Działania dezinformacyjne, prowadzone przez inne kraje, są zaplanowane jako element wojny hybrydowej i obejmują oddziaływanie w przestrzeni informacyjnej, cyberprzestrzeni.

W Polsce podatność na takie operacje jest bardzo duża, ze względu na niski poziom zaufania społecznego i silną polaryzację. (...)

Dezinformacja, manipulacja, propaganda to jedne z najważniejszych narzędzi walki informacyjnej, jednej z płaszczyzn **wojny hybrydowej**. Pierwszym krokiem jest autorefleksja.

Ogromne znaczenie mają **kompetencje medialne i odporność informacyjna** obywateli.

Paulina Piasecka

Jak wspierać budowanie ODPORNOŚCI PSYCHICZNEJ?

ANALIZA ŹRÓDEŁ

- Kto jest autorem treści?
- Skąd pochodzi informacja?
- Jakie są intencje nadawcy?
- Czy inne źródła to potwierdzają?

ROZPOZNAWANIE DEEPPFAKE'ÓW

Nie technicznie, a **poznawczo**:

- Czy treść wywołuje bardzo silne emocje?
- Czy nagranie jest „zbyt idealne” lub zbyt gładkie?
- Czy nagle pojawia się kontrowersyjna wypowiedź znanej osoby?
- Czy ktoś zyskuje na tym, że w to uwierzysz?

Zanim uwierzysz, zastanów się, kto na tym korzysta.

HIGIENA INFORMACYJNA

- ogranicz źródła do kilku zaufanych,
- nie udostępniaj niczego „na emocjach”,
- wprowadź zasadę 10 sekund: zatrzymaj się → pomyśl → sprawdź.

BUDOWANIE ODPORNOŚCI EMOCJONALNEJ

Ucz:

- identyfikacji stanu emocjonalnego („co czuję, gdy to widzę?”),
- rozróżniania myśli od faktów,
- kontroli uwagi,
- regulacji pobudzenia.

CO ROBIĆ?

- omawiać konkretne **przypadki** dezinformacji i deepfake'ów.
- wprowadzić „analizę źródeł” jako **stały element**.
- pracować na **kontrastach**: „oryginał vs manipulacja”.
- pokazywać, jak wyglądają **kampanie informacyjne**
- współtworzyć **szkolne zasady** korzystania z materiałów generowanych przez AI.
- budować **kulturę pytań**: „Skąd to wiesz?” „Jak możesz to sprawdzić?”
- wzmacniać **własną odporność**, bo uczniowie uczą się przede wszystkim przez **modelowanie**.

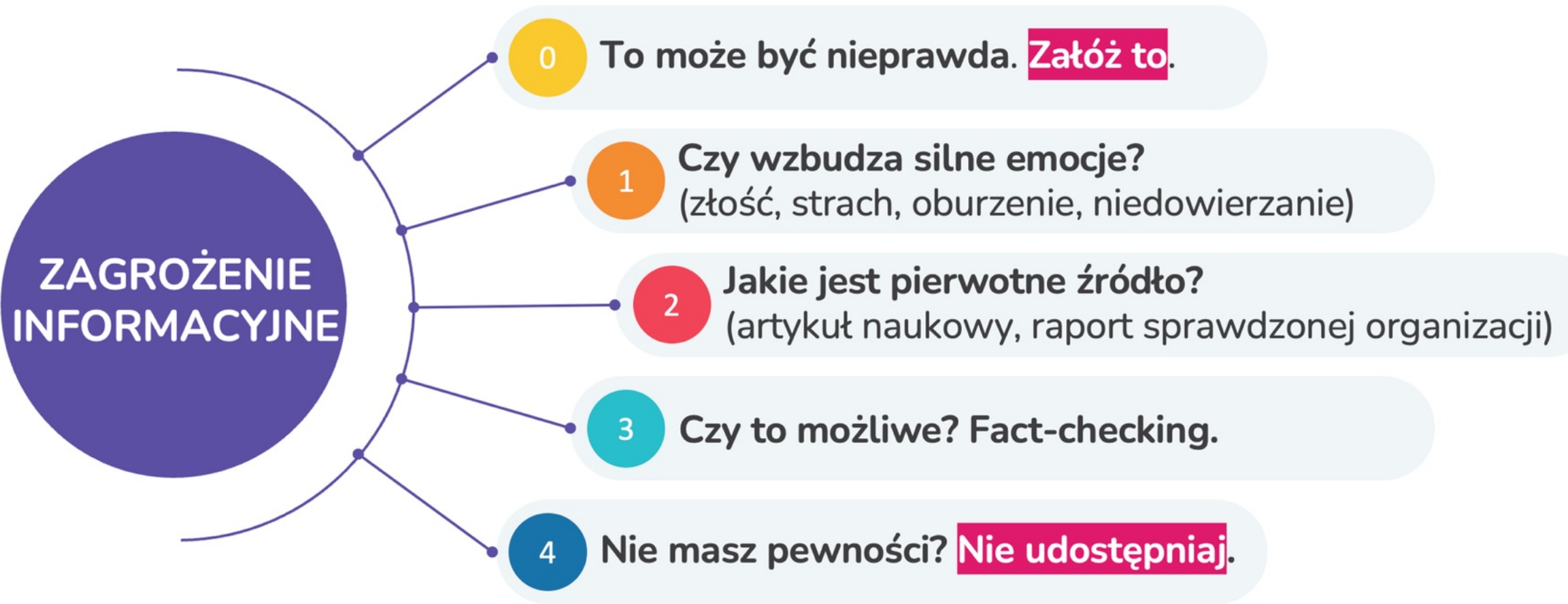
EDUKACJA NA RZECZ ODPORNOŚCI PSYCHICZNEJ

Odporność poznawcza na dezinformację - umiejętność **rozpoznawania** manipulacji, deepfake'ów, wojny narracyjnej.

Regulacja stresu i odporność psychiczna - wojna informacyjna i realne zagrożenie działają przez **emocje**.

Wiedza o mózgu i mechanizmach lęku - lęk jest naturalny, ale nie może prowadzić do paraliżu. Musimy **wiedzieć, jak działa mózg** pod wpływem strachu, jak podejmować decyzje mimo zagrożenia.

Kompetencje społeczne oparte na współpracy - izolacja i atomizacja społeczeństwa to idealne środowisko dla przeciwnika. Edukacja musi budować zdolność do tworzenia **wspólnoty i zaufania** (fundament odporności narodowej).



Pisanie

w epocie post-AI

Pisanie hybrydowe

człowiek-AI staje się normą.

Próba ustalenia, gdzie kończy się człowiek, a gdzie zaczyna AI, jest bezcelowa i daremna.

Pisanie przestało być procesem liniowym.

To proces iteracyjny, metapoznawczy i sterowany oceną.

Daję polecenie i **oceniam**, co dostaję.

Analizuję (Jaka jest struktura? Gdzie są luki? Czy AI używa emocjonalnych odwołań? Gdzie halucynuje?)

Na podstawie tego, co mam, **stosuję nową strategię** w kolejnym prompcie.

Znowu **oceniam** (Czy zbliżam się do tego, czego potrzebuję?)

Buduję osobistą bazę wiedzy o tym, co sprawia, że osiągam porządany efekt.

AI ma wspierać, a nie zastępować, ma być narzędziem wzmacniającym ludzką refleksję, nie jej substytutem.

1. **Start with Human:** badania muszą zaczynać się od ludzkich pytań, ciekawości, intuicji i kontekstu. AI nie potrafi zrozumieć znaczenia problemów, priorytetów społecznych czy etycznych. To człowiek nadaje sens i kierunek badaniom.
2. **Work with AI, not through AI:** AI może wspierać analizę danych, język, kodowanie jakościowe, prognozowanie, ale człowiek sprawuje kontrolę i ponosi odpowiedzialność. Giray ostrzega przed „AI shaming” (potępianiem osób, które jawnie używają AI mimo transparentności i etycznego podejścia).
3. **End with Human:** końcowa interpretacja, sens, wartości, etyka, publikacja i odpowiedzialność za treść muszą należeć do badacza. AI może poprawiać tekst, ale nie może odpowiadać za znaczenie, logikę ani prawdę naukową.

Strategia "Human–AI–Human"

Człowiek inicjuje, prowadzi i kończy proces badawczy.

TechTrends
<https://doi.org/10.1007/s11528-025-01132-7>

AECT

BRIEF COMMUNICATION

When Using AI in Scientific Research: Start with Human, End with Human

Louie Giray¹

Received: 14 April 2025 / Accepted: 19 August 2025
© Association for Educational Communications & Technology 2025

Abstract

Researchers are increasingly encouraged to adopt AI tools, yet there is still a lack of practical guidance on how to use AI without compromising the human essence of scholarly work. This paper addresses this gap by proposing a “start with human, end with human” approach to AI integration in research. At the core of this proposal is the Human–AI–Human cycle, a framework that maintains human agency at both ends of the research process. The cycle begins with the human scholar identifying problems, and formulating research questions and design, which all are grounded in human values and expertise. AI is then employed to assist with data processing, simulating, analyzing, and suggesting. Finally, the process returns to the human scholar, who handles the interpretation, sense-making, ethical review, peer dialogue, and public dissemination, all of which require human agency and accountability. This balanced approach ensures that AI amplifies rather than diminishes human scholar capabilities. It encourages scholars to embrace AI not as a threat, but as a tool that, when used wisely, can enhance, not erode, the soul of scientific inquiry. The implications are significant for research ethics, scientific practices, and the evolving definition of academic labor in the age of AI.

Keywords Artificial intelligence · Academic integrity · Writing · Research ethics · Human-AI collaboration · Scientific research

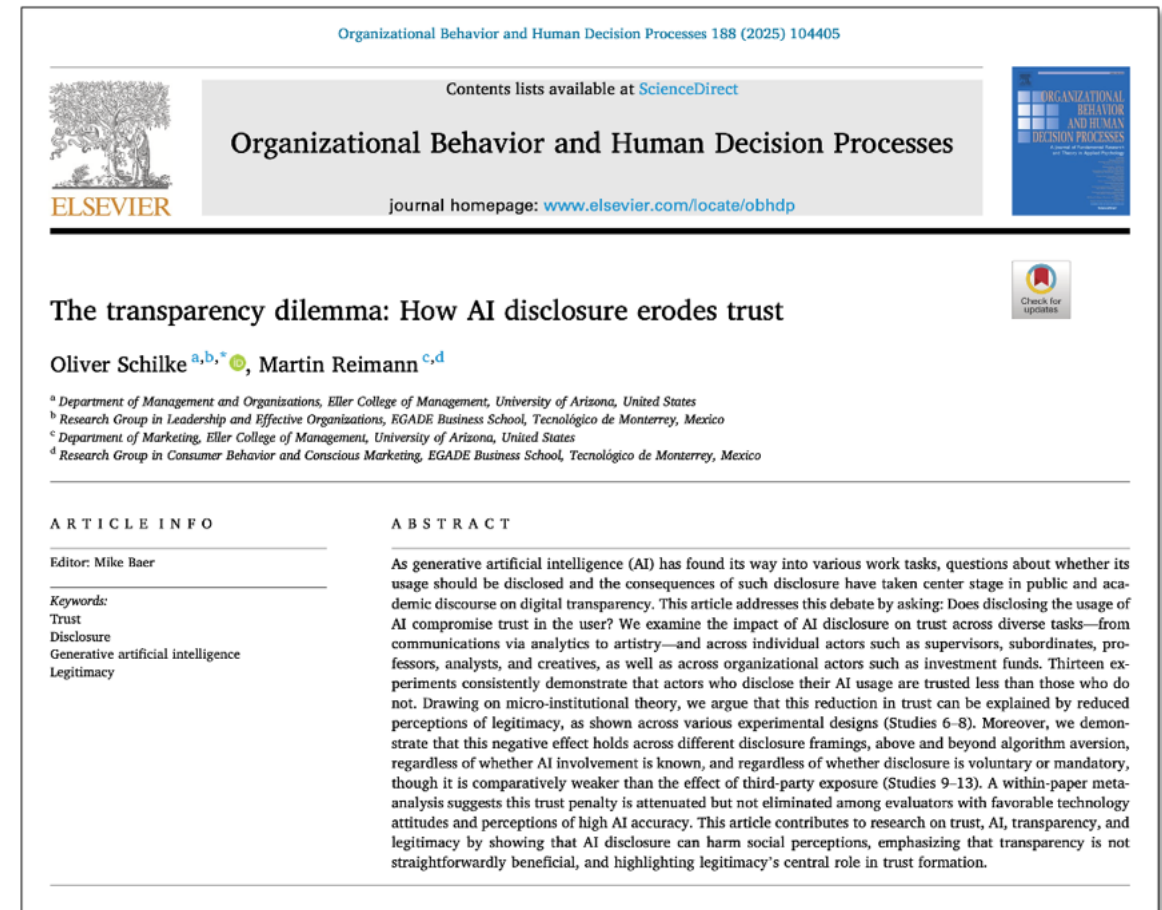
Osoby ujawniające korzystanie z narzędzi AI w wykonaniu zadania otrzymują **niższy poziom zaufania** niż w przypadku, gdy użycia AI nie ujawniono.

Efekt ten utrzymuje się niezależnie od:

- formy ujawnienia (dobrowolnej czy wymuszonej)
- tego, czy odbiorca wiedział o użyciu AI wcześniej czy nie
- poziomu dokładności narzędzia AI

Ludzie uznają, że użycie AI może być mniej „czyste”, mniej autentyczne.

„Przejrzystość dotycząca udziału AI wywołuje wrażenie mniejszej autentyczności i zaangażowania osobistego, co podważa zaufanie interpersonalne.”



6 zasad postplagiatu: pisanie w epoce AI

Sarah Elaine Eaton

23 February 2023

Rozdział w: "Plagiarism in Higher Education: Tackling Tough Topics in Academic Integrity" (2021)

Hybrydowe pisanie (człowiek-AI) stanie się normą

Tworzenie tekstów wspólnie z AI staje się powszechne i wkrótce będzie standardem. Próby ustalenia, gdzie kończy się wkład człowieka, a zaczyna praca AI, są bezcelowe.

Kreatywność człowieka zyskuje, nie traci

AI nie zagraża naszej twórczości, może ją wręcz pobudzać. Ludzie potrafią inspirować AI, a AI może inspirować nas. Nasza zdolność do myślenia, wyobraźni i tworzenia pozostaje niewyczerpana, nawet w erze technologii.

Bariery językowe będą zanikać

Pierwszy język zacznie mieć coraz mniejsze znaczenie, gdy narzędzia staną się dostępne dla ludzi do wzajemnego zrozumienia w niezliczonych językach.



Człowiek może oddać część kontroli, ale nie odpowiedzialność

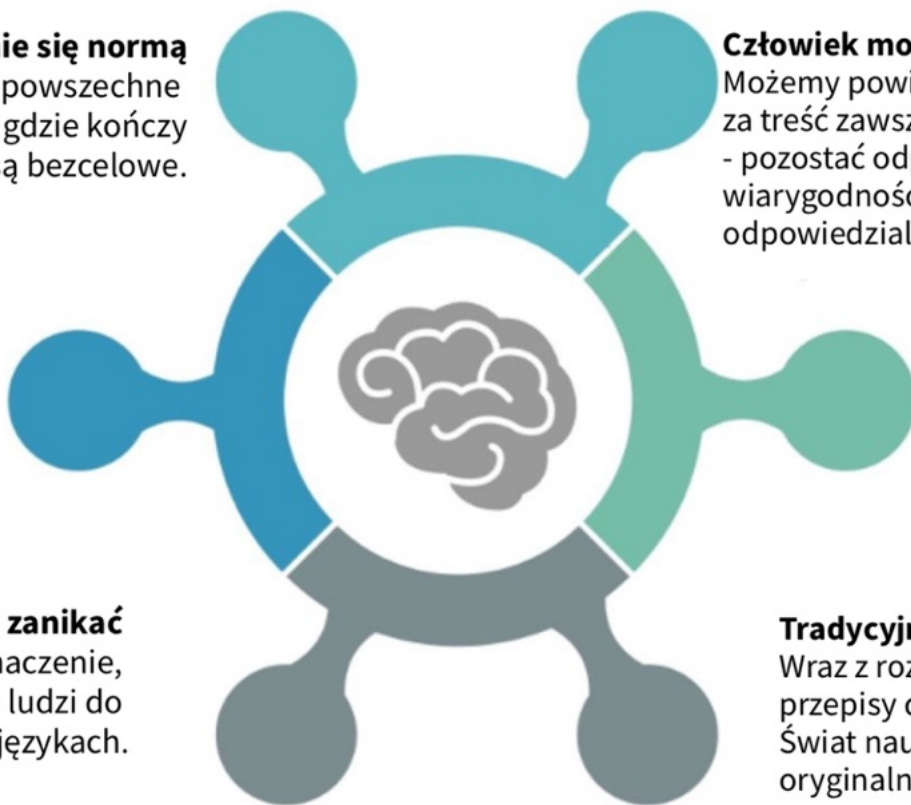
Możemy powierzać AI część procesu pisania, ale odpowiedzialność za treść zawsze pozostaje po naszej stronie. Ludzie mogą - i muszą - pozostać odpowiedzialni za sprawdzanie faktów, rzetelność, wiarygodność i etyczne użycie narzędzi. Ludzie są również odpowiedzialni za to, jak opracowywane są narzędzia AI.

Uznanie autorstwa nadal ma znaczenie

Docenianie źródeł, cytowanie, odwoływanie się do innych - to fundament etyki akademickiej, który nie traci aktualności. Uczymy się od nauczycieli, mentorów i społeczności. Wskazywanie inspiracji i uznanie wkładu innych to nadal - i zawsze będzie - ważna umiejętność.

Tradycyjne definicje plagiatu przestają wystarczać

Wraz z rozwojem AI nie chodzi o to, by na nowo pisać przepisy dotyczące plagiatu, lecz by wyjść poza nie. Świat nauki i edukacji musi zrozumieć, że reguły oceny oryginalności wymagają nowego języka i nowych zasad.





Academic Integrity in a Postplagiarism Era:

Prioritizing Pedagogy instead of Punishment

Sarah Elaine Eaton, PhD
Brock University
May 5, 2025



Dr. Sarah Eaton, Academic Integrity in a Postplagiarism Era

JAK WESPRZEĆ STUDENTÓW?

(w czytaniu, pisaniu, głębokim myśleniu, refleksji)

Jak wspierasz w czytaniu?

- w sylabusie **ograniczona** liczba pozycji, wskazane **rozdziały**
- wysyłany krótki materiał do **samodzielnego** zapoznania się
- polecaj książki **popularno-naukowe**
- oceniał **ustne** wypowiedzi
- **ucz, jak czytać** artykuły naukowe (wybierz **fragment**, zadaj **pytania** do niego, daj pola na odpowiedzi pisane **ręcznie**)

Zamiast "Przeczytaj artykuł"

- wypisz **trudne pojęcia** i sprawdź je w parze/na forum
- zaznacz i skomentuj: „Czego **nie rozumiem?**”, „Co mnie **zaskoczyło?**”
- dyskusja: co autor miał na myśli, a **co ja zrozumiałem/am?**
- znajdź fragment, który budzi twój **sprzeciw** i uzasadnij dlaczego
- jakie **pytania** z tego tekstu warto zadać AI?
- przygotuj **3 pytania** na bazie tego tekstu do dyskusji grupowej
- wykonaj **mapę myśli** / notatkę Cornella
- streść jeden akapit w **15 słowach**
- napisz 10 **słów kluczowych** do artykułu
- **sparafrazuj** sens akapitu

Porównaj z odpowiedzią AI. Przeanalizuj różnice, przedyskutuj, wyciągnij wnioski.

Rutyny myślenia

Proste, powtarzalne struktury myślowe, które wspierają **analizę, ocenę i rozwiązywanie problemów.**

Pomagają w uporządkowaniu procesu myślowego, ułatwiając **krytyczną ocenę** informacji i **podejmowanie decyzji.**

Zestaw Narzędzi Rutyny Myślenia Projekt Zero, University of Harvard



Rutyna myślenia to zestaw pytań lub krótka sekwencja kroków używanych do wspierania krytycznego myślenia.

O głębokim czytaniu

”

Samotność, powolność, ciągła uwaga nie są już stanami domyślnymi, ale **aktami oporu**.

Te warunki ulegają erozji, możliwość spokojnej lektury jest coraz rzadsza. Głęboka lektura pozwalała na zamieszkanie w obcych umysłach, zmaganie się ze sprzecznymi ideami. Czytać znaczy mówić: "Twoje słowa mają znaczenie. Twoja złożoność ma znaczenie. Twoja myśl ma znaczenie." Chodzi o to, by zatrzymać impuls do odpowiadania, przewijania.
I w tym zawieszeniu pojawia się: empatia, refleksja, przemiana.

Algorytm nie dba o znaczenie, a o ruch, ma na celu nie pogłębienie naszej percepcji, a doprecyzowanie naszych preferencji.

Carl Hendrick

PRZYKŁADY WYKORZYSTANIA AI PRZEZ STUDENTÓW I DOKTORANTÓW

Dobre praktyki.

AI jako "partner i krytyk"

"Większość kadry akademickiej nie ma doświadczenia w korzystaniu z AI, podobnie jak my, studenci, ale wiemy, że wielu z nas eksperymentuje i bada, w jaki sposób może być przydatna w uczeniu się podczas naszych studiów."

Dobre praktyki w korzystaniu z AI przez studentów i doktorantów.

1. Lepsze zrozumienie pojęć
2. Nauka rachunkowości i finansów
3. Przygotowanie do egzaminów
4. Od bajtów do mózgu: sojusz edukacyjny
5. Zwiększenie produktywności akademickiej i zawodowej
6. Planowanie zadań
7. Zwiększanie klarowności
8. Mój nowy przyjaciel
9. Część mojego codziennego życia studenckiego
10. Pomędzy wątpliwością a wsparciem
11. Nauka pływania w morzu liczb
12. Rozwijanie pewności siebie łączenie kropek
13. Tworzenie rusztowania dla innowacji
14. W stronę samodzielności
15. Od ciekawości do kreatywności
16. Lepsze zrozumienie teorii psychologicznych
17. Tworzenie scenariuszy
18. Nie droga na skróty, lecz partner do nauki
19. Motywacja do rozmowy
20. Partner i krytyk
21. Partner myślowy i asystent badawczy
22. Nauka programowania
23. GenAI i sztuka utrzymania motocykla
24. Złota rączka, mistrz wszystkiego?
25. Czasami czuję się bardziej komfortowo z GenAI
26. Współtworzenie wiedzy
27. Myślenie z AI



LEARNING WITH AI

A Student Edited Collection

KAŻDY UCZY SIĘ INACZEJ

Personalizacja uczenia się

Jak AI wspiera personalizację?

- **Adaptacyjne ścieżki:** tutor AI wspiera w procesie uczenia się, analizuje odpowiedzi i proponuje kolejne zadania dopasowane do poziomu, pracuje metodą sokratejską, daje feedback itd.
- **Wielostylowe treści:** AI przekształca materiał w różne formaty pozwalając uczniowi wybrać najlepszy dla siebie sposób.
- **Dynamiczne dopasowanie języka i poziomu trudności:** AI potrafi przeformułować ten sam tekst na poziom podstawowy, średnio zaawansowany i akademicki.
- **Refleksja metapoznawcza:** coach AI może zadawać pytania „Jak byś to wytłumaczył koledze?”, „Który krok był dla ciebie najtrudniejszy?”.
- **Generowanie scenariuszy:** uczeń może prosić AI o zadania kontekstowe, co zwiększa poczucie sensu i wyboru. Symulacje i role-play: AI wciela się w rolę partnera do rozmowy (nauka języków).
- **Personalizowane plany sesji powtórek (spaced repetition):** AI proponuje powtórki w coraz dłuższych odstępach czasu.

Jak AI wspiera personalizację?

- **Feedback dopasowany do potrzeb ucznia:** niektórzy wolą krótkie wskazówki, inni długie wyjaśnienia z przykładami.
- **Wzmacnianie mocnych stron:** AI analizuje, w jakich typach zadań uczeń radzi sobie najlepiej (np. analiza danych, praca kreatywna) i częściej podsuwa takie formy pracy.
- **Dostosowanie modalności:** „Wyjaśnij mi to na schemacie”, „Zwizualizuj to”, „Wytłumacz głosowo”.
- **Scaffolding i stopniowe zdejmowanie podpór:** AI początkowo prowadzi ucznia krok po kroku, a potem celowo daje coraz mniej wskazówek.
- **Personalizacja kontekstu kulturowego i zainteresowań:** AI osadza zadanie w kontekście zainteresowań (np. trasy żółwi morskich, misja kosmiczna, science fiction, islamska kaligrafia)
- **Wsparcie refleksji i autorefleksji:** AI kończy sesję pytaniami: „Co dziś było dla ciebie najbardziej wartościowe?” „Jak wykorzystasz tę wiedzę w praktyce?” „Co chcesz powtórzyć jutro?”

Cechy tekstu generowanego przez AI

Wskazówki, jak wykryć nieujawnione treści generowane przez AI

Katalog powszechnych wzorców tekstu generowanego przez AI, specyficznych dla Wikipedii.

1. Nieuzasadniony nacisk na **symbolikę** i znaczenie ("odgrywa znaczącą rolę", "jest symbolem")
2. **Powierzchowne analizy**.
3. **Promocyjny** i pozytywnie naładowany język.
4. Dydaktyczne, redakcyjne **zastrzeżenia** ("Należy jednak pamiętać...").
5. **Streszczenia** sekcji ("Podsumowując...")
6. **Konkluzje** mówiące o wyzwaniach i perspektywach na przyszłość.
7. **Negatywne równoległości** ("Nie tylko ... ale ...").
8. **Zarys negatywów** ("Nie kariera, nie praca, nie trwałe znaczenie - tylko moment algorytmiczny.")
9. **Niejasne przypisywanie** opinii ("Badania pokazują;...")
10. Reguła **trzech** elementów.

To tylko wybrane wskazówki, pełna lista dostępna jest na stronie https://en.wikipedia.org/wiki/Wikipedia:Signs_of_AI_writing



Ocenianie

w epocie post-AI

Potrzebujemy na nowo pomyśleć
o ocenianiu:

ocena jako filtr

czy narzędzie rozwoju?



Oceniasz efekt czy proces?

Tradycyjny model oceniania skoncentrowany na **efekcie końcowym** traci swoją funkcję rozwojową i diagnostyczną (i nie jest to nowy problem).

Modele językowe potrafią wygenerować poprawne odpowiedzi, ale dają iluzję mistrzostwa i tworzą **rusztowanie zastępujące proces głębokiego myślenia**.

Im bardziej mierzymy wynik, tym mniej widzimy, jak przebiega proces uczenia się, z czym się mierzy osoba ucząca się i kim się staje w trakcie tego procesu.

Ocenić proces **stawania się**.

Jak osoba ucząca się myślała? Jak się zmieniła?

Ocenianie jako **czytanie śladów przemiany**.

Potrzebujemy **nowego sposobu myślenia o ocenianiu** efektów uczenia się, który powinien obejmować elementy:

- **droga dojścia do celu**
- **refleksja**
- **etyczne użycie AI**

jako części rozwoju poznawczego.

droga dojścia do celu
refleksja
etyczne użycie AI



- AI log
- Samoocena (opis, skala)
- Self-report z procesu
- Dziennik refleksji (w tym etyczne użycie)
- AI: kiedy pomagała, kiedy przeszkadzała?
- Do czego nie użyłem/am AI i dlaczego?

AI log

Krótki i treściwy: 8-15 zdań, ma pokazywać proces myślowy (nie historię użycia narzędzia).

1. Do czego użyłem AI?

- plan zadania / szkic / kod / pytania pomocnicze (dokładny cel użycia)

2. Konkretnie prompty

- 2-5 najważniejszych promptów
- pokazanie, co dokładnie AI dostała jako wejście

3. Fragmenty odpowiedzi AI (lub ich streszczenie)

- tylko te elementy, które realnie wykorzystałem
- bez przepisywania całych stron

4. Czy opisałem poprawki, które wprowadziłem?

Najważniejszy element:

- co poprawiłem
- dlaczego poprawiłem
- w czym AI się pomyliła
- czego AI nie uwzględniła
- co musiałem doprecyzować

5. Czy pokazałem miejsca, w których AI pomogła / przeszkodziła?

- pomogła: np. wygenerowała kod, strukturyzowała plan
- przeszkodziła: np. halucynowała wyniki, myliła wzory, proponowała nierealne założenia

6. Opis co zrobiłem samodzielnie

- obliczenia
- interpretacja wyników
- analiza danych
- argumentacja
- decyzje projektowe

7. Czy AI log pozwala prowadzącemu ocenić zrozumienie?

- czy z logu wynika, że student wie, co robi?
- czy potrafi odróżnić poprawną sugestię AI od błędnej?

OPISOWA

SAMOOOCENA jako element procesu oceny

Refleksja nad procesem uczenia się:

1. **Co było dla Ciebie największym wyzwaniem** w tym zadaniu/projekcie?
Dlaczego?
2. **Z jakich strategii korzystałeś/aś**, aby poradzić sobie z tym wyzwaniem?
3. **Czy korzystałeś/aś z AI?** Jeśli tak, to jak? Co zrobiła AI, a co Ty?
4. **Co Cię zaskoczyło** lub czego się nauczyłeś/aś podczas pracy nad tym zadaniem?
5. **Co zrobił(a)byś inaczej**, gdybyś zaczynał/a od nowa?

SKALA

SAMOOOCENA jako element procesu oceny

1. **Zaangażowanie w proces uczenia się** (1 – brak zaangażowania, 5 – intensywne i świadome zaangażowanie)
2. **Samodzielność myślenia i decyzji** (1 – głównie kopiowanie/przyjmowanie gotowców, 5 – samodzielna analiza i wybory)
3. **Umiejętność refleksji nad własnym uczeniem się** (1 – nie zastanawiałem/am się nad tym, 5 – głęboka refleksja)
4. **Etyczne i świadome użycie AI** (1 – brak świadomości / ukrywanie, 5 – przejrzyste, krytyczne wykorzystanie)
5. **Gotowość do dalszej poprawy** (1 – brak, 5 – silna chęć rozwoju)

Dziennik refleksji

jako narzędzie oceniania

Jak doświadczenie pracy z AI kształtuje moje myślenie, mój rozwój, mój proces uczenia się?

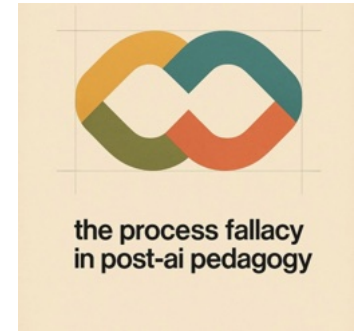
Refleksja to świadome zatrzymanie się, przemyślenie własnych doświadczeń, działań i decyzji. Buduje krytyczne myślenie, wspiera tworzenie pamięci, wzmacnia poczucie sprawczości, motywację i dobrostan, rozwija metapoznanie.

A może **rytuał** robienia namysłu nad sposobem pracy z narzędziami AI przez studentów?

Refleksyjny dziennik (papierowy lub cyfrowy), w którym studenci zapisują swoje przemyślenia, refleksje, wahania, frustracje, pytania, czyli swoje myśli i emocje, które towarzyszą im w trakcie procesu uczenia się i pracy z AI.

ale!

uważaj na nadmierne "dokumentowanie myślenia"!



Na biurokratyzowanie refleksji.

Znaczenie może powstawać w sposób, który opiera się dokumentacji.

Tworzenie obejmuje momenty, których żadne kryteria oceniania nie są w stanie uchwycić.

Może produkt jako ciągłe powstawanie? Żywy artefakt?

Celem nie jest rejestrowanie etapów myślenia, ale projektowanie zadań, które wspierają myślenie i rozwój.

Za: Owen Matson (zmienione)

Refleksja do każdego zadania jako nowy element (samo)oceny:

Kiedy AI była pomocna?

Kiedy AI nie była pomocna?

(utrudniała, wydłużała pracę)

Zmiana warsztatu dydaktycznego

Od "mierzenia wiedzy" → do **wspierania rozwoju kompetencji** poznawczych i etycznych



Od wskazywania błędów → do **tutoringu i refleksyjnego feedbacku**



Od "mierzenia wiedzy" → do **wspierania rozwoju kompetencji** poznawczych i etycznych

- Zamiast sprawdzać, czy student zapamiętał, sprawdzaj jak **myśli**, jak **analizuje**, jak **argumentuje**.
- **Ocenianie nie kończy procesu**, jest jego częścią (pętla feedbackowa, wersje robocze/iteracje).
- Etos pracy poznawczej: **nie „zrób dobrze”, lecz „zrób świadomie i odpowiedzialnie”**.

Co robić na zajęciach?

- **Zadania z otwartą strukturą poznawczą** (np. „Znajdź trzy sprzeczne interpretacje tego problemu i oceń ich trafność”).
- **Zadania z wyborem strategii**: student opisuje, jak rozwiązał problem (nie tylko wynik).
- **Refleksyjne podsumowania**: Co było trudne? Czego się nauczyłem/am? Co bym zrobił/a inaczej?
- **Dyskusje etyczne o korzystaniu z AI**: nie jako kontrola, ale rozmowa o granicach, wartościach i celach.

Od wskazywania błędów → do **tutoringu i refleksyjnego feedbacku**

- Nie „odsiewaj”, ale **pomagaj się rozwijać** (Ty jako przewodnik).
- Feedback nie jest „komentarzem do pracy”, lecz **narzędziem transformacji myślenia**.

Co robić na zajęciach?

- **Praca na wersjach roboczych** (iteracyjna): np. projekt 1, 2, 3 z komentarzami i samooceną.
- **"Rubyki" oceniania wspierające rozwój** z kryteriami takimi jak „metapoznanie”, „strategia uczenia się”, „etyczne decyzje”, „refleksyjność”.
- Metoda „**feedforward**”: zamiast wyłącznie oceny tego, co było, wskazanie, co i jak poprawić w kolejnych działaniach.

Refleksja nad moim procesem oceniania

Czy mój sposób oceniania wspiera rozwój studentów?

Oceniam nie tylko efekt końcowy, ale również proces dochodzenia do efektu (proces uczenia się).

Udzielam studentom informacji zwrotnej, która zawiera konkretne wskazówki rozwojowe.

Pozwalam studentom pracować nad wersjami roboczymi i poprawkami swoich prac.

Stosuję kryteria oceny, które uwzględniają refleksję, strategie uczenia się i zaangażowanie poznawcze.

Zachęcam studentów do samooceny i autorefleksji nad procesem uczenia się.

Ocenianie traktuję jako część procesu rozwoju, a nie jedynie sposób selekcji.

Uwzględniam rozwój kompetencji etycznych i poznawczych, nie tylko treści merytorycznych.

Umożliwiam studentom ujawnienie, w jaki sposób korzystali z narzędzi AI (zachęcam).

Moje zadania są tak zaprojektowane, że wymagają myślenia, analizy, decyzji, a nie tylko odtwarzania wiedzy.

Daję studentom przestrzeń na rozmowę o kryteriach, celach i sensie oceniania.

Strategie oceniania wspierające rozwój i odporne na AI

1. **Ocena formacyjna zamiast sumatywnej.** Praca na wersjach roboczych. Feedback jako dialog, nie jako wyrok. Ujawnianie toku rozumowania (metapoznanie, komentarze, autoanaliza).
2. **Ocenianie procesów poznawczych i strategii uczenia się.** Dzienniki uczenia się, refleksje, mapy myślenia. Opis użytych strategii oraz trudności i ich pokonania.
3. **Wspólne projektowanie kryteriów z studentami.** Student wie, za co jest oceniany i współtworzy standardy jakości.
4. **Ocenianie współpracy i argumentacji, nie tylko treści.** Dyskusje, debaty, współtworzone treści.
5. **Ujawniamy korzystanie z AI i uczymy odpowiedzialności.** Zadania z refleksją: "Co zrobiłeś/aś z pomocą AI, a co sam/a?". Zadania wymagające krytycznej analizy odpowiedzi AI (co działa, co nie, gdzie były błędy).

Zaktualizuj swój sylabus

Krytyczna umiejętność korzystania z GenAI:

- zdolność do krytycznej analizy i angażowania się w systemy generatywnej sztucznej inteligencji poprzez zrozumienie ich podstaw technicznych, wyzwań etycznych oraz wpływu na proces uczenia się,
- świadomość ograniczeń, uprzedzeń oraz szerszych skutków społecznych, środowiskowych i gospodarczych.

SYLABUS

Efekty uczenia się

„Nabycie wiedzy i umiejętności w zakresie...”

Propozycja zmiany: Rozwój umiejętności refleksyjnych i metapoznawczych. Świadome i etyczne korzystanie z narzędzi AI w procesie uczenia się. Rozwijanie samodzielności poznawczej i krytycznego myślenia.

Metody dydaktyczne

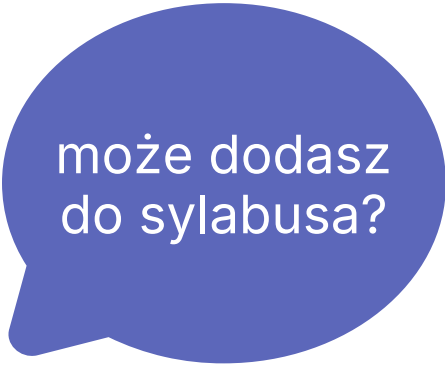
Dodaj: Praca na wersjach roboczych (iteracyjne uczenie się). Dzienniki refleksji, autoanaliza procesu uczenia się. Dyskusje wokół odpowiedzi generowanych przez AI. Zadania porównawcze: „z AI” vs „bez AI” z omówieniem strategii.

Metody i kryteria oceniania

Egzamin końcowy, projekt zaliczeniowy

Propozycja zmiany: Portfolio dokumentujące proces uczenia się. Refleksyjny esej: „Czego się nauczyłem/em i jak?”. Krytyczna analiza narzędzi AI i strategii ich użycia. Rubryki oceny obejmujące: proces, myślenie, krytyczną autorefleksję, etyczne decyzje.

SYLABUS



może dodasz
do sylabusa?

„W ramach przedmiotu studenci będą zachęceni do eksplorowania potencjału i ograniczeń AI w kontekście własnego uczenia się. Ocenie podlegać będzie nie tylko efekt końcowy, ale również dochodzenia do rozwiązania, rozważania natury etycznej oraz refleksja metapoznawcza.”

WSKAZÓWKI WYKORZYSTANIA AI W ZADANIU

AI ASSESSMENT SCALE (AIAS)

BEZ UŻYCIA AI

Praca (zadanie) jest realizowana w całości bez wsparcia AI w kontrolowanym środowisku, co gwarantuje, że studenci polegają wyłącznie na swojej wiedzy, rozumieniu i umiejętnościach.
Nie używaj AI na żadnym etapie realizacji zadania, tworzenia pracy.
Wykaż się swoją własną wiedzą i umiejętnościami.

PLANOWANIE Z AI

AI może być używane do działań przed wykonaniem zadania, takich jak burza mózgów, tworzenie konspektu i wstępne rozeznanie (research). Cel: efektywne wykorzystanie AI do planowania, syntezy i generowania pomysłów, jednak ocena powinna uwzględniać zdolność do samodzielnego rozwijania i doskonalenia tych pomysłów.
Możesz korzystać z AI w celu planowania, rozwijania pomysłów, prowadzenia wstępnych badań źródłowych.
Twoja praca powinna pokazywać, w jaki sposób rozwijałeś/aś i doskonalisz/aś te pomysły.

WSPÓŁPRACA Z AI

AI może być wykorzystywane do pomocy w realizacji zadania, w tym do generowania pomysłów, pisania, otrzymywania informacji zwrotnych i doskonalenia pracy. Student powinien krytycznie oceniać i modyfikować sugestie AI, pokazując rozumienie procesu i zaangażowanie w proces twórczy.
Możesz korzystać z AI, by wykonać konkretne zadanie, np. pisanie tekstu, doskonalenie i ocenianie swojej pracy.
Musisz krytycznie oceniać i modyfikować wszelkie treści wygenerowane przez AI, z których korzystasz.

PEŁNE WSPARCIE AI

AI może być wykorzystywane do realizacji dowolnych elementów zadania (pracy), a studenci instruuja AI tak, aby zrealizować swój cel. Ocena pracy może wymagać interakcji z AI lub analizy drogi dojścia studenta do celu, by właściwie ocenić jego sposób pracy z AI i dojścia do rozwiązania problemu.
Możesz szeroko korzystać z AI podczas pracy według własnych potrzeb lub zgodnie z wytycznymi dotyczącymi realizacji zadania, także w trakcie egzaminu.
Skoncentruj się na kierowaniu AI, aby osiągnąć swoje cele, jednocześnie wykazując się krytycznym myśleniem.

POZNAWANIE MOŻLIWOŚCI AI

AI jest wykorzystywane kreatywnie do rozwiązywania problemów, generowania nowych, opracowywania innowacyjnych rozwiązań. Studenci i nauczyciele wspólnie ustalają ocenę pracy (ocena uzgodniona), aby odkrywać unikalne zastosowania AI w danym obszarze.
Korzystaj z AI w sposób kreatywny, aby wykonać zadanie, rozwiązać problem, być może współprojektując nowe podejścia z nauczycielem lub innymi osobami.

POZIOM 1: NO AI

BEZ UŻYCIA AI

Opis:

Student wykonuje całe zadanie samodzielnie: obliczenia, analizę, interpretację, decyzje, wnioski, kod, raport.

AI nie jest używana na żadnym etapie.

Kiedy stosować:

- gdy kluczowe jest rozumienie podstaw,
- w laboratoriach (interpretacja danych),
- w zadaniach, gdzie efekt końcowy musi odzwierciedlać myślenie studenta,
- przy sprawdzaniu wiedzy i samodzielności poznawczej.

Rola nauczyciela:

Zaprojektować zadania tak, aby nie mogły być rozwiązane przez AI (np. oparte na danych z zajęć, kontekście lokalnym).

POZIOM 2: AI FOR PLANNING

PLANOWANIE Z AI

Opis:

AI pomaga studentowi przygotować się do zadania:

- checklisty,
- opis kroków,
- podpowiedzi organizacyjne,
- wyjaśnienie ogólnych zasad (bez danych z zadania).

AI nie wykonuje obliczeń, interpretacji, kodu, rozwiązania.

Kiedy stosować:

- gdy chcesz wesprzeć studentów w organizacji pracy, bez wpływania na treść,
- gdy zadanie obliczeniowe/eksperymentalne wymaga 100% samodzielności,
- gdy zadanie jest odporne na AI, ale proces można usprawnić.

Rola nauczyciela:

Wyjaśnić, że AI nie może „liczyć ani analizować”, tylko pomaga uporządkować proces.

POZIOM 3: AI FOR DRAFTING + CRITIQUE

WSPÓŁPRACA Z AI

Opis:

AI generuje **szkic** rozwiązania:

- szkic kodu,
- szkic struktury raportu,
- szkic analizy,
- pomysł na podejście.

Student **musi poprawić**, uzupełnić, zweryfikować i uzasadnić zmiany.

Refleksja i krytyka są kluczowe.

Kiedy stosować:

- zadania projektowe, analityczne, programistyczne,
- gdy chcesz rozwijać myślenie krytyczne wobec AI,
- gdy szkic nie jest wartością samą w sobie (np. pisanie esejów, struktur kodu).

Rola nauczyciela:

Wymagać AI logu i jawnego wskazania „co AI dała, a co student poprawił”.

POZIOM 4: AI FULL USE WITH DOCUMENTATION

PEŁNE WSPARCIE AI

Opis:

AI może wykonać **duże fragmenty** pracy:

- generować kod,
- proponować architekturę modelu,
- podpowiadać rozwiązania numeryczne,
- przygotować prototyp lub szkic raportu.

Student musi zrozumieć, wyjaśnić i ocenić wygenerowaną treść oraz podjąć własne decyzje projektowe.

Kiedy stosować:

- w zadaniach, gdzie AI jest naturalnym narzędziem,
- gdy celem jest praca z narzędziami, nie „samodzielne wynajdywanie koła”,

Rola nauczyciela:

Ocenić zrozumienie: „wyjaśnij własnymi słowami”, „co poprawiłeś?”

POZIOM 5: AI CO-CREATION

POZNAWANIE
MOŻLIWOŚCI AI

Opis:

Student i AI **współtworzą** rozwiązanie: iteracje, poprawki, wersje alternatywne, optymalizacja.

Proces jest najważniejszy.

Student bada kompromisy, wybiera opcje, uzasadnia decyzje.

AI jest partnerskim narzędziem.

Kiedy stosować:

- projekty otwarte, zespołowe, twórcze,
- zaawansowane kursy projektowe (IT, budownictwo, ML),
- prace, gdzie jest wiele równorzędnych rozwiązań.

Rola nauczyciela:

Patrzeć na proces iteracyjny, nie efekt:

„Pokaż mi trzy iteracje i powiedz, dlaczego wybrałeś akurat tę”.

Rozszerzona wersja skali

Poziom 6: AI exploration (poziom meta)

To poziom badawczy.

- student bada halucynacje, błędy, biasy,
- testuje limity modelu,
- nie rozwiązuje zadania „z pomocą AI”, ale analizuje AI jako obiekt badań.

Odkrywanie granic i błędów AI:

„kiedy AI się myli?”, „dlaczego halucynuje?”, „jak ją skorygować?”.

To poziom, który świetnie nadaje się do zadań projektowych i refleksyjnych.

Ćwiczenie

Do każdego zdania (opis sytuacji) zaproponuj poziom skali AIAS:

1. „Poprosiłem AI o listę kroków do rozwiązania zadania.”
2. „AI rozwiązała zadanie, a ja nie bardzo wiem, jak to działa.”
3. „AI zrobiła szkic raportu, a ja poprawiłem dane i interpretację.”
4. „Sam zrobiłem obliczenia; AI pomogła mi upewnić się, że nie zrobiłem błędu.”
5. „AI wygenerowała trzy propozycje, a ja eksperymentuję i porównuję je.”
6. „Nie użyłem AI, muszę to rozumieć sam.”
7. „AI napisała kod, ja go modyfikuję i tłumaczę, co działa, a co nie.”
8. „AI wygenerowała błędne wnioski, poprawiłem je i zapisałem dlaczego.”
9. „AI proponowała rozwiązanie, które ja usprawniam.”
10. „AI wygenerowała wiele opcji, a ja odkrywam ich ograniczenia.”
11. „AI pomogła mi poukładać raport, ale treść jest moja.”
12. „Zabrałem zdjęcia z laboratorium i sam opisałem to, co widzę.”

Rozwiązanie

POZIOM 1: NO AI

Student wykonuje zadanie w całości samodzielnie.

6. „Nie użyłem AI, muszę to rozumieć sam.”

12. „Zabrałem zdjęcia z laboratorium i sam opisałem to, co widzę.”

POZIOM 2: AI FOR PLANNING

AI pomaga tylko w organizacji procesu, nie w rozwiązaniu.

1. „Poprosiłem AI o listę kroków do rozwiązania zadania.”

11. „AI pomogła mi poukładać raport, ale treść jest moja.”

POZIOM 3: AI FOR DRAFTING + CRITIQUE

AI tworzy szkic, student poprawia, koryguje, krytycznie analizuje.

3. „AI zrobiła szkic raportu, a ja poprawiłem dane i interpretację.”

8. „AI wygenerowała błędne wnioski, poprawiłem je i napisałem dlaczego.”

POZIOM 4: AI FULL USE WITH DOCUMENTATION

AI wykonuje dużą część pracy, student musi zrozumieć, wyjaśnić, udokumentować.

4. „Sam zrobiłem obliczenia; AI pomogła mi upewnić się, że nie zrobiłem błędu.”

7. „AI napisała kod, ja go modyfikuję i tłumaczę, co działa, a co nie.”

POZIOM 5: AI CO-CREATION

Student i AI współtworzą rozwiązanie iteracyjnie, twórczo.

5. „AI wygenerowała trzy propozycje, a ja eksperymentuję i porównuję je.”

9. „AI zaproponowała rozwiązanie, które ja usprawniam.”

POZIOM 6: AI EXPLORATION

Student bada działanie AI, jej ograniczenia, błędy, halucynacje.

2. „AI rozwiązała zadanie, a ja nie bardzo wiem, jak to działa.”

10. „AI wygenerowała wiele opcji, a ja odkrywam ich ograniczenia.”

Analiza i przebudowa instrukcji i kryteriów oceniania wybranego zadania

Przykład 1

Przykład 1: DRGANIA TŁUMIONE (WILIŚ, WETI)

Temat: Analiza układu równań różniczkowych opisującego drgania tłumione.

Treść zadania:

Rozważ układ drgań tłumionych opisany równaniem:

$$m \cdot x'' + c \cdot x' + k \cdot x = 0$$

gdzie:

$$m = 2 \text{ kg}, c = 6 \text{ Ns/m}, k = 10 \text{ N/m}.$$

1. Wyznacz wartość współczynnika tłumienia i określ, czy układ jest podtłumiony, krytycznie tłumiony, czy przetłumiony.
2. Wyprowadź analityczne rozwiązanie równania ruchu.
3. Zinterpretuj fizycznie zachowanie układu.
4. Zasymuluj przebieg $x(t)$ dla pierwszych 10 sekund.

Cel dydaktyczny:

rozumienie równań ruchu, klasyfikacji drgań tłumionych, powiązania matematyki z fizyką.

Co jest oceniane?

Poprawność obliczeń i rozwiązania analitycznego.

Umiejętność interpretacji fizycznej przebiegu $x(t)$.

Jakość symulacji i krytyczne wykorzystanie narzędzi.

Do decyzji:

- Co student musi policzyć sam?
- Gdzie symulacja może być wsparta AI (np. kod w Pythonie)?
- Czy interpretację można robić z pomocą AI, ale z krytyczną weryfikacją?
- Jak udokumentować użycie AI (AI log)?



Przykład 1: DRGANIA TŁUMIONE (WILIŚ, WETI)

Co student musi policzyć sam?

- współczynnik tłumienia ξ
- klasyfikację układu (pod-/krytycznie-/przetłumiony)
- analityczne rozwiązanie równania ruchu
- interpretację fizyczną zachowania układu

Gdzie AI może wspierać obliczenia?

- wygenerowanie kodu symulacyjnego (Python/MATLAB)
- tworzenie wykresów $x(t)$
- zapis równania ruchu w formie znormalizowanej
- wskazówki dotyczące solverów ODE

Czy interpretację można robić z pomocą AI?

- tak, ale student musi skonfrontować opis z przebiegiem
- AI często halucynuje zachowania tłumienia (np. zbyt wolny/za szybki zanik amplitudy)
- interpretacja musi odnosić się do własnej symulacji

Jak udokumentować użycie AI (AI log)?

- jakie prompty zastosowano
- którą część kodu wygenerowała AI
- co student poprawił i dlaczego
- gdzie AI pomogła, a gdzie była myląca

Przykład 1: DRGANIA TŁUMIONE (WILIŚ, WETI)

Cel dydaktyczny:

rozumienie równań ruchu, klasyfikacji drgań tłumionych, powiązania matematyki z fizyką.

Co jest oceniane?

Poprawność obliczeń i rozwiązania analitycznego.

Umiejętność interpretacji fizycznej przebiegu $x(t)$.

Jakość symulacji i krytyczne wykorzystanie narzędzi.

AI w zadaniu (podsumowanie):

AIAS 3: AI for drafting + critique. AI do: kodu symulacyjnego, technicznego wsparcia matematycznego.

Bez AI: obliczenia, rozwiązanie analityczne, interpretacja.

AI log obowiązkowy.

AI w zadaniu (wersja pełna)



Przykład 1: DRGANIA TŁUMIONE (WILIŚ, WETI)

KROK 1. Określenie poziomu AIAS

Poziom 3: AI for drafting + critique

KROK 2. Uzasadnienie wybór poziomu AIAS

1. Współczynnik tłumienia, klasyfikacja (pod-/krytycznie-/przetłumiony) i rozwiązanie analityczne wymagają samodzielnej pracy matematycznej. AI nie może tego zrobić za studenta, jeśli chcę rozwijać rozumienie równań różniczkowych.
2. AI może pomóc technicznie: w zapisie równań, doborze solvera ODE, wygenerowaniu kodu do symulacji, to uzasadnia poziom 3 (drafting + critique), a nie 1-2.
3. Interpretacja przebiegu $x(t)$ musi być oparta na własnych wykresach i parametrach, których AI nie widzi, tu chcę zachować No AI.
4. Poziom 4 byłby zbyt wysoki: AI mogłaby przejąć całą symulację i interpretację, co osłabiłoby efekt dydaktyczny.
5. Poziom 3 pozwala ćwiczyć krytyczne sprawdzanie kodu/obliczeń AI, a jednocześnie zachowuje rdzeń poznawczy zadania jako w pełni ludzki.

KROK 3. Zapis do sylabusu (zasady AI)

Poziom AIAS: 3 (AI for drafting + critique)

Dozwolone:

1. Użycie AI do wygenerowania szkicu kodu symulacyjnego (Python/MATLAB) dla danego równania.
2. Prośba o pomoc w zapisie równań w postaci standardowej ODE.
3. Konsultacja dotycząca użycia solverów (jak ustawić krok czasowy, przedział symulacji).
4. Korekta językowa krótkiej interpretacji (styl, klarowność), ale nie treść merytoryczna.

Zabronione:

1. Generowanie pełnego rozwiązania analitycznego równania ruchu przez AI.
2. Prośzenie AI o klasyfikację układu (pod-/krytycznie-/przetłumiony) na podstawie danych.
3. Generowanie gotowej interpretacji fizycznej zachowania układu i wniosków.
4. Oddanie raportu wygenerowanego w całości lub w większości przez AI.

Wymagana dokumentacja (AI log):

- do jakich elementów użyto AI (np. kod, zapis ODE),
- przykładowy prompt,
- co AI wygenerowała,
- jakie poprawki wprowadził student i dlaczego,
- które części rozwiązania są całkowicie samodzielne.

Przykład 1: DRGANIA TŁUMIONE (WILIŚ, WETI)**KROK 4. Zadanie po przebudowie**

Zadanie 1.0 (oryginał): rozwiąż, zasymuluj, zinterpretuj.

Zadanie 1.2 (po przebudowie, z polityką AI):

1. Samodzielnie oblicz współczynnik tłumienia, częstotliwość własną i zaklasyfikuj typ układu.
2. Samodzielnie wyprowadź rozwiązanie analityczne równania ruchu.
3. Możesz użyć AI do wygenerowania kodu do symulacji $x(t)$ (10 s), pamiętaj o AI logu.
4. Narysuj wykres $x(t)$ i opisz własnymi słowami zachowanie układu (amplituda, częstotliwość, zanikanie).
5. Dołącz AI log, w którym pokazujesz, jak AI pomogła (lub przeszkodziła) i co zrobiłeś sam.

DETEKTORY AI

Co wyniki badań mówią o ich wiarygodności?

"Moi studenci oszukują"

Jednak nie tu leży sedno problemu. I to **nie jest nowy problem,** jedynie skala jest większa.

To kryzys założeń programów nauczania, sposobów oceniania, to **kryzys obecnej pedagogiki.**

A detektory nie rozwiążą naszych problemów.
Czas na "AI pedagogy".

Generatywne systemy wykrywania AI **nie są obecnie i prawdopodobnie nie będą** w stanie w przewidywalnej przyszłości wykrywać treści generowanych przez AI z wystarczającą dokładnością, aby można było na nich polegać w sprawach takich, jak oszustwa akademickie.

Co mówią wyniki badań na temat skuteczności "detektorów" AI



Analiza skuteczności detektorów AI

Perkins et al. 2024

Analiza skuteczności detektorów AI (m.in. GPTZero, Turnitin, DetectGPT) wobec tekstów poddanych prostym modyfikacjom (parafraza, wstawianie synonimów, zmiana szyku zdań).

Wyniki:

- dla czystych tekstów AI skuteczność detekcji wynosiła ok. **39,5%**.
- Po zastosowaniu prostych technik „maskowania” spadła do **17,4%**.

Wnioski:

Detektory są bardzo podatne na obejścia i **nie mogą być traktowane jako narzędzia do decyzji** o wysokiej wadze (np. oszustwa akademickie).



Springer Open

Search  [Get pu](#)

[International Journal of Educational Technology in Higher Education](#)

[About](#) [Articles](#) [Article Collections](#) [Submission Guidelines](#) [Videos](#) [Infographics](#) [Submit manuscript !\[\]\(ff0f641c578c5123e9861807558a635e_img.jpg\)](#)

Research article | [Open access](#) | Published: 09 September 2024

Simple techniques to bypass GenAI text detectors: implications for inclusive education

[Mike Perkins](#) , [Jasper Roe](#), [Binh H. Vu](#), [Darius Postma](#), [Don Hickerson](#), [James McGaughran](#) & [Huy Q. Khuat](#)

[International Journal of Educational Technology in Higher Education](#) **21**, Article number: 53 (2024) | [Cite this article](#)

480 Accesses | **217** Altmetric | [Metrics](#)

Abstract

This study investigates the efficacy of six major Generative AI (GenAI) text detectors when confronted with machine-generated content modified to evade detection (n = 805). We compare these detectors to assess their reliability in identifying AI-generated text in educational settings, where they are increasingly used to address academic integrity concerns. Results show significant reductions in detector accuracy (17.4%) when faced with simple techniques to manipulate the AI generated content. The varying performances of GenAI tools and detectors indicate they cannot currently be recommended for determining academic integrity violations due to accuracy limitations and the potential for false accusation which undermines inclusive and fair assessment practices. However, these tools may support learning and academic integrity when used non-punitively. This study aims to guide educators and institutions in the critical implementation of AI text detectors in higher education, highlighting the importance of exploring alternatives to maintain inclusivity in the face of emerging technologies.

Analiza skuteczności detektorów AI

Weber-Wulff et al. 2023

Cel badań

Skuteczności detektorów AI: 12 narzędzi publicznych (m.in. GPTZero, Writer, Sapling) i 2 komercyjne (Turnitin, PlagiarismCheck).

Wyniki

- detektory wykazują niską dokładność, częściej klasyfikują teksty AI jako „ludzkie” niż odwrotnie
- skuteczność spada drastycznie, gdy teksty są lekko parafrazowane lub „zamaskowane”
- narzędzia komercyjne (Turnitin, PlagiarismCheck) radziły sobie niewiele lepiej niż darmowe

Wnioski

Detektory nie nadają się do stosowania jako dowód w sprawach o wysokiej wadze (np. oszustwa akademickie)

International Journal for Educational Integrity

Home About [Articles](#) Submission Guidelines

[Submit manuscript](#) 

Original article | [Open access](#) | Published: 25 December 2023

Testing of detection tools for AI-generated text

[Debora Weber-Wulff](#), [Alla Anohina-Naumeca](#), [Sonja Bjelobaba](#) , [Tomáš Foltýnek](#), [Jean Guerrero-Dib](#), [Olumide Popoola](#), [Petr Šigut](#) & [Lorna Waddington](#)

[International Journal for Educational Integrity](#) **19**, Article number: 26 (2023) | [Cite this article](#)

83k Accesses | **240** Citations | **768** Altmetric | [Metrics](#)

Abstract

Recent advances in generative pre-trained transformer large language models have emphasised the potential risks of unfair use of artificial intelligence (AI) generated content in an academic environment and intensified efforts in searching for solutions to detect such content. The paper examines the general functionality of detection tools for AI-generated text and evaluates them based on accuracy and error type analysis. Specifically, the study seeks to answer research questions about whether existing detection tools can reliably differentiate between human-written text and ChatGPT-generated text, and whether machine translation and content obfuscation techniques affect the detection of AI-generated text. The research covers 12 publicly available tools and two commercial systems (Turnitin and PlagiarismCheck) that are widely used in the academic setting. The researchers conclude that the available detection tools are neither accurate nor reliable and have a main bias towards classifying the

Analiza skuteczności detektorów AI

Elkhatat et al. 2023

Cel badań

Ocena skuteczności najpopularniejszych narzędzi do wykrywania treści generowanych przez AI (m.in. OpenAI Text Classifier, GPTZero, Copyleaks, Writer, CrossPlag).

Wyniki

- skuteczność narzędzi silnie zróżnicowana (lepsza przy GPT-3.5, wyraźnie słabsza przy GPT-4)
- występowały liczne fałszywe alarmy: teksty ludzkie oznaczano jako „AI”.
- narzędzia często dawały niejednoznaczne klasyfikacje („prawdopodobnie AI”), co zmniejszało ich wartość praktyczną

Wnioski

Detektory AI nie są wystarczająco rzetelne, aby opierać na nich decyzje akademickie lub zawodowe.

Original article | [Open access](#) | Published: 01 September 2023

Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text

[Ahmed M. Elkhatat](#) ✉, [Khaled Elsaid](#) & [Saeed Almeer](#)

[International Journal for Educational Integrity](#) **19**, Article number: 17 (2023) | [Cite this article](#)

122k Accesses | 217 Citations | 171 Altmetric | [Metrics](#)

Abstract

The proliferation of artificial intelligence (AI)-generated content, particularly from models like ChatGPT, presents potential challenges to academic integrity and raises concerns about plagiarism. This study investigates the capabilities of various AI content detection tools in discerning human and AI-authored content. Fifteen paragraphs each from ChatGPT Models 3.5 and 4 on the topic of cooling towers in the engineering process and five human-written control responses were generated for evaluation. AI content detection tools developed by OpenAI, Writer, Copyleaks, GPTZero, and CrossPlag were used to evaluate these paragraphs. Findings reveal that the AI detection tools were more accurate in identifying content generated by GPT 3.5 than GPT 4. However, when applied to human-written control responses, the tools exhibited inconsistencies, producing false positives and uncertain

Analiza skuteczności detektorów AI

Popkov & Barrett 2024

Cel badań

Ocena skuteczności popularnych detektorów AI na tekstach akademickich (100 artykułów pisanych przez ludzi oraz 200 tekstów generowanych przez modele językowe (ChatGPT, Claude). Porównano narzędzi darmowych i komercyjnych.

Wyniki

- zarówno fałszywie dodatnie (teksty ludzkie oznaczane jako AI), jak i fałszywie ujemne
- niektóre detektory radziły sobie lepiej z prostymi esejami, ale zawodziły przy złożonych artykułach naukowych
- różnice w skuteczności były znaczne między narzędziami, ale żadne nie osiągnęło poziomu wiarygodnego

Wnioski

Detektory AI są zawodne w kontekście naukowym i nie mogą być traktowane jako podstawa do stwierdzania plagiatu czy oszustwa.



Advanced

Save

Email

> [Account Res.](#) 2024 Mar 22:1-17. doi: 10.1080/08989621.2024.2331757. Online ahead of print.

AI vs academia: Experimental study on AI text detectors' accuracy in behavioral health academic writing

Andrey A Popkov ^{1 2}, Tyson S Barrett ¹

Affiliations + expand

PMID: 38516933 DOI: [10.1080/08989621.2024.2331757](https://doi.org/10.1080/08989621.2024.2331757)

Abstract

Artificial Intelligence (AI) language models continue to expand in both access and capability. As these models have evolved, the number of academic journals in medicine and healthcare which have explored policies regarding AI-generated text has increased. The implementation of such policies requires accurate AI detection tools. Inaccurate detectors risk unnecessary penalties for human authors and/or may compromise the effective enforcement of guidelines against AI-generated content. Yet, the accuracy of AI text detection tools in identifying human-written versus AI-generated content has been found to vary across published studies. This experimental study used a sample of behavioral health publications and found problematic false positive and false negative rates from both free and paid AI detection tools. The study assessed 100 research articles

Teksty human-origin i detektory AI

Saha & Feizi 2025

Cel badań

Ocena skuteczności detektorów tekstów pisanych przez ludzi, które zostały jedynie „wygładzone” przez AI (np. poprawa gramatyki, stylu, płynności).

Wyniki

- wiele narzędzi klasyfikowało teksty jako AI-generated, mimo że treść i argumentacja pochodziły od człowieka.
- granica między tym, co „ludzkie” a „AI” staje się bardzo płynna w przypadku pisania hybrydowego.
- wysoki poziom wyników fałszywie dodatnich podważa praktyczne zastosowanie detektorów

Wnioski

Detektory nie są w stanie rozróżnić subtelnych przypadków współpracy człowieka z AI. Autorzy wskazują na konieczność nowych standardów oceny autentyczności.

Almost AI, Almost Human: The Challenge of Detecting AI-Polished Writing

Shoumik Saha
University of Maryland
College Park, USA
smksaha@umd.edu

Soheil Feizi
University of Maryland
College Park, USA
sfeizi@umd.edu

Abstract

The growing use of large language models (LLMs) for text generation has led to widespread concerns about AI-generated content detection. However, an overlooked challenge is AI-polished text, where human-written content undergoes subtle refinements using AI tools. This raises a critical question: should minimally polished text be classified as AI-generated? Such classification can lead to false plagiarism accusations and misleading claims about AI prevalence in online content. In this study, we systematically evaluate *twelve* state-of-the-art AI-text detectors using our AI-Polished-Text Evaluation (APT-Eval) dataset, which contains 15K samples refined at varying AI-involvement levels. Our findings reveal that detectors frequently flag even minimally polished text as AI-generated, struggle to differentiate between degrees of AI involvement, and exhibit biases against older and smaller models. These limitations highlight the urgent need for more nuanced detection methodologies.

📄 github.com/ShoumikSaha/ai-polished-text

📄 huggingface.co/datasets/smksaha/apt-eval

1 Introduction

The rapid advancement of LLMs has enabled AI to generate highly fluent, human-like text, raising

Additionally, reports suggesting that a significant share of online content is AI-generated – such as analyses indicating that over 40% of Medium posts are likely AI-written – often overlook the nuance of AI-polished text, which detection tools may misclassify.^{2,3} These sweeping claims risk misrepresenting the actual extent of AI involvement, leading to misleading statistics and misplaced skepticism about human authorship. Motivated by these issues, our study systematically examines how AI-text detectors respond to AI-polished text and whether their classifications are both accurate and fair.

To investigate this issue, we introduce the **AI-Polished-Text Evaluation (APT-Eval)** dataset of size **15K**, which systematically examines how AI-text detectors respond to varying degrees of AI involvement in human writing. Our dataset is built from pre-existing human-written samples that are refined using different LLMs, such as GPT-4o (OpenAI, 2023), Llama3-70B (Dubey et al., 2024), DeepSeek-V3 (Liu et al., 2024), etc., applying degree and percentage based modifications. This allows us to assess how detectors respond to minor and major AI polishing. We analyze the classification accuracy, false positive rates, and domain-specific sensitivities of **12 state-of-the-art detectors** spanning model-based, metric-based

Stanowisko ACM **U.S. Technology Policy Committee** dotyczące zasad tworzenia i używania systemów do wykrywania treści generowanych przez AI.

Rosnąca dostępność i użycie AI (teksty, obrazy, audio) wywołały **zapotrzebowanie na systemy**, które potrafią wiarygodnie wykrywać, czy dane treści zostały w całości lub częściowo stworzone przez AI.

Oczekiwania różnych sektorów:

Edukacja: uczelnie chciałyby wiedzieć, czy prace studentów były przygotowane z pomocą AI.

Pracodawcy: chcą wykrywać użycie AI w aplikacjach o pracę.

Media: chcą odróżniać komentarze ludzkie od tych generowanych przez chatboty.

Administracja publiczna: potrzebuje rozróżniać autentyczne listy i komentarze od treści generowanych algorytmicznie.

Obecnie dostępne technologie detekcji **nie są wystarczająco wiarygodne**, by można było podejmować na ich podstawie decyzje krytyczne, wpływające na życie i karierę ludzi (np. karać studentów, odrzucać kandydatów do pracy czy uznawać oficjalne dokumenty za fałszywe).



STATEMENT ON PRINCIPLES FOR THE DEVELOPMENT AND USE OF SYSTEMS TO DETECT GENERATIVE AI CONTENT

The ACM US Technology Policy Committee (USTPC)¹ notes that the dramatic increase in the availability, proliferation, and use of generative artificial intelligence technology in all sectors of society has created concomitant growing demand for systems that can reliably detect when a document, image, or audio file contains information produced in whole or in part by a generative AI system. Specifically, for example:

- Educational institutions want systems that can reliably detect when college applications and student assignments were created with the assistance of generative AI systems;
- Employers want systems that can detect the use of generative AI in job applications;
- Media companies want generative AI systems so that they can distinguish human comments on their articles from responses generated by chatbots; and
- Government agencies need to tell human letters and comments from responses that were algorithmically generated.

*Demand for such systems, however, is no measure of their accuracy² or fairness.³ Indeed, the Committee finds and cautions that **no such presently available detection technology is sufficiently reliable on which to exclusively base critical, potentially life- and career-altering decisions** in the contexts and use cases cited above, or any other. Accordingly, while AI detection systems may provide useful preliminary assessments, their outputs should never be accepted as proof of AI-generated content.*

Sprawdź możliwości detektorów AI.

The image shows a side-by-side comparison of a text editor and the GPTZero AI detection interface. On the left, a text editor window titled 'Mózg nie jest kamerą, tylko p...' contains a paragraph of text. On the right, the GPTZero 'Basic scan' interface shows the same text being analyzed. A green circle with the word 'human' indicates the detection result. A probability breakdown shows 0% AI generated, 0% Mixed, and 100% Human. Below this, 'Advanced Sentence Scanning' and 'Your most human sentences' are visible.

Mózg nie jest kamerą, tylko p...

Mózg nie jest kamerą, tylko plasteliną.
Nie zapisuje rzeczy „jak leci”, tylko je ugniata, skleja, przetwarza. Każde wspomnienie to rzeźba – im więcej razy do niej wracasz, im więcej dotykasz, poprawiasz, tym twardsza i bardziej trwała. Badania o tzw. konsolidacji pamięci pokazują, że ślady pamięciowe stabilizują się w korze mózgowej po kilkukrotnym powtarzaniu i odświeżaniu (Dudai, 2012, Neuron; Squire & Wixted, 2011, Annual Review of Neuroscience).
Co to oznacza dla Ciebie?
Samo „czytanie w kółko” to jakbyś ugniatał plastelinę tylko raz – zostanie miękka i łatwo się rozpadnie.
Testowanie siebie (quizey, flashcards, recall) to jak wyciąganie rzeźby z szafy i sprawdzanie, czy jeszcze stoi. Każde wyjęcie ją wzmacnia (Roediger & Butler, 2011, Trends in Cognitive Sciences).

Basic scan

⚠ We partially support Polski.
We're still working on this language, so the results may be imperfect. Request your language [here](#).

GPTZero AI Detection Model 3.2m

human

We are highly confident this text is entirely human

Probability breakdown ⓘ

0% AI generated 0% Mixed 100% Human

View AI Sentences AI Words

Advanced Sentence Scanning

Understand how each sentence impacts AI probabilities and with interpretable reports (American Federation Teacher Certified)

Your most human sentences

Badania o tzw.
Każde wyjęcie ją wzmacnia (Roediger & Butler, 2011...

Scan to update results
779 characters 116 words

Scan →

To jest tekst w całości wygenerowany przez AI (chatGPT) i wklejony do GPTzero.

To jest wynik (całkowicie nieprawdziwy)

GPTzero



NEWS

AI-Detectors Biased Against Non-Native English Writers

DATE

MAY 15, 2023

TOPICS

DESIGN, HUMAN-COMPUTER INTERACTION

NATURAL LANGUAGE PROCESSING

MACHINE LEARNING

Don't put faith in detectors that are "unreliable and easily gamed," says scholar.

In the wake of the high-profile launch of ChatGPT, no fewer than seven developers or companies have countered with AI detectors. That is, AI they say is able to tell when content was written by another AI. These new algorithms are pitched to educators, journalists, and others as tools to flag cheating, plagiarism, and mis- or disinformation.

It's all very meta, but according to [a new paper](#) from Stanford scholars, there's just one (very big) problem: The detectors are not particularly reliable. Worse yet, they are especially unreliable when the real author (a human) is not a native English speaker.



SHARE



[Advice and Guidance](#)

AI Detection and assessment – an update for 2025

 By [Michael Webb](#) [24 June 2025](#) [No Comments](#)

poprzednie artykuły

-  Webb M. 2023. **AI writing detectors – concepts and considerations.** National Centre for AI
-  Webb M. 2023. **AI Detection – Latest Recommendations.** National Centre for AI

DALSZE WSPARCIE

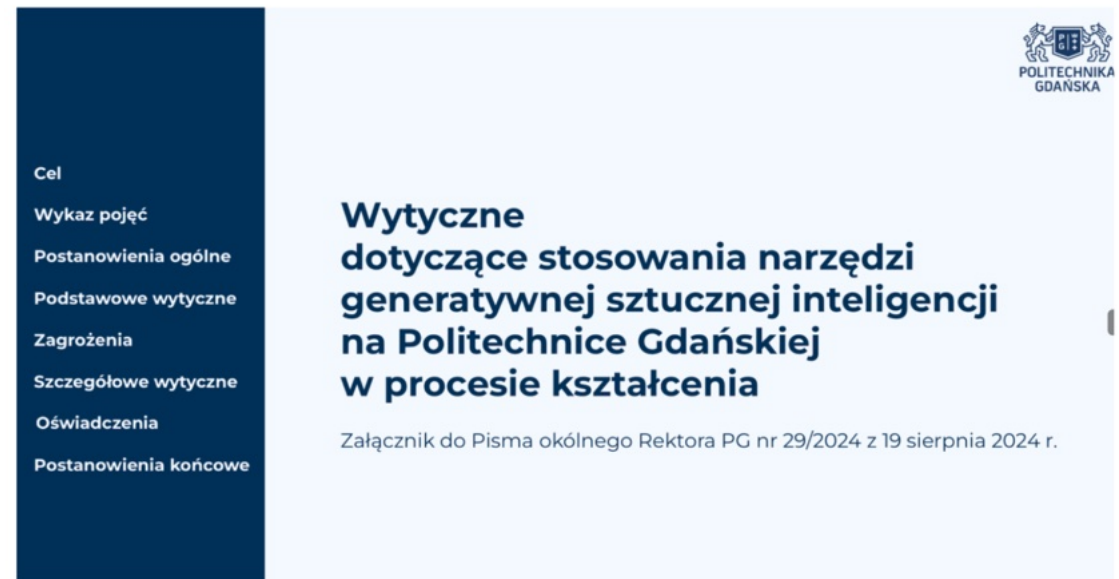
Co oferuje Ci uczelnia?



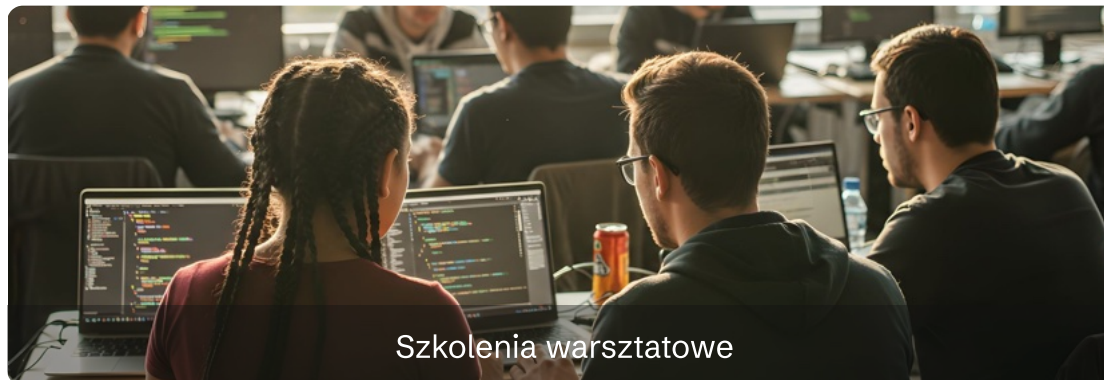
Strona www AI na PG



Ambasadorzy metodyki AI



Zasady wykorzystania AI w pracach dyplomowych

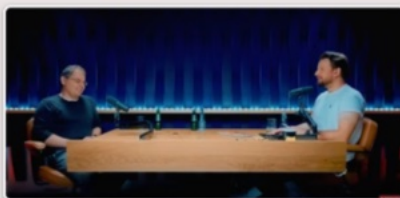


Szkolenia warsztatowe



Materiały wideo

więcej...



Aleksander Mądry



Andrzej Dragan



Jacek Dukaj



Dragan, Dukaj



UNESCO Guidance AI in EDU



UNESCO AI for teachers



UNESCO AI for students



AI Pedagogy by Harvard



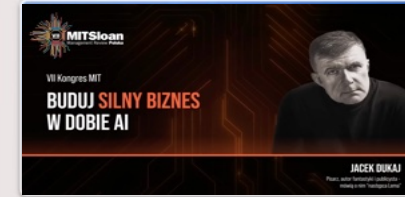
Mary Meeker's AI Report 2025



Dragan, Duch, Gawron



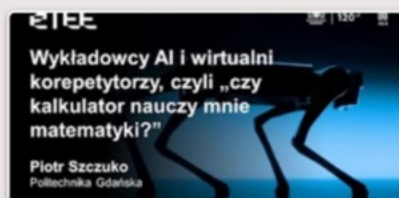
Dragan, Dukaj



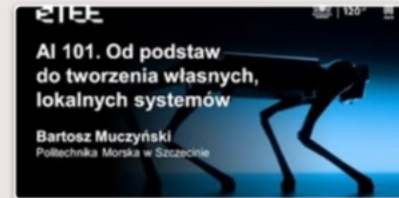
Jacek Dukaj



Ethan's Mollick blog



Piotr Szczuko eTEE



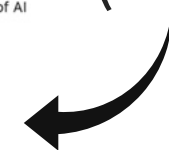
Bartosz Muczyński eTEE



Tomasz Jankowski Ideatorium

Raporty Wytyczne Badania Trendy Analizy

(folder z plikami PDF)



UNESCO. AI competency framework for teachers.pdf



UNESCO AI competencies for students.pdf



UNESCO Guidance for generative AI in education and research.pdf



UNESCO. AI and education. Guidance for policy makers.pdf



UNESCO. K-12 AI curricula.pdf



UNESCO April 2023. Education 2030 ChatGPT...er education .pd



UNESCO Understanding the impact of AI skill...pment 2021.pdf



UNESCO 2022 Higher education figures at a glance.pdf



Generative_AI_and_ChatGPT_in_Higher_Education.pdf



McKinsey 2024 he-state-of-ai-in-early-2024-final.pdf



McKinsey 2024 mckinsey-technology-tre...utlook-2024.pdf



McKinsey. 2023. The state of AI in 2023.pdf



WEF Shaping the Future of Learning 2024.pdf



AI-education report USA 06.2023.pdf



COLORADO Roadmap for AI in K12 Education.pdf



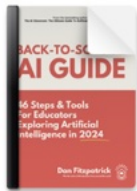
UK Use Cases for Generative AI in Education.pdf



Overview Unlocking the Value of AI in Higher Ed.pdf



Brown 2023. AI generation. How universities prepare students.pd



Fitzpatrick. 2024. Back to school AI guide.pdf



Digital Education Council Global AI Student Survey 2024.pdf



TREDNS 2024_Artificial-Intelligence.pdf



10 KEY points AI competency for teachers UNESCO.jpeg



K-12 rubrics AISG 2024.jpeg



peoples-guide-ai 2018.pdf



AI for Education (Intermediate)



Instruktor: [Jeanne Beatrix Law](#)



AI in Education: Leveraging ChatGPT for Teaching



Instruktor: [Ethan Mollick](#)



Specjalizacja Generative AI for Educators & Teachers

Innovate in your classroom and school. Master techniques for using ChatGPT to enhance rather than harm education



Instruktor: [Dr. Jules White](#) **Najlepszy instruktor**

A może jakiś kurs e-learningowy?

coursera



Accelerate Your Learning with ChatGPT



Instruktorzy: [Dr. Jules White](#) +1 więcej **Najlepszy instruktor**



Generatywna sztuczna inteligencja dla każdego



Instruktor: [Andrew Ng](#) **Najlepszy instruktor**

EDUxAI dla studentów

a także uczniów i ich rodziców

Joanna Mytnik

Przewodnik dla
uczniów
i studentów
w uczeniu się
ze wsparciem AI

Q TL;DR



Jak **się** uczyć efektywnie?

Garść informacji o tym, jak działa Twój mózg + zestaw narzędzi i metod efektywnego uczenia się.

Joanna Mytnik

Interaktywny
przewodnik dla
uczniów
i studentów

Q TL;DR