Kryzys autorstwa w epoce sztucznej inteligencji: funkcje tekstu, granice kreatywności i odpowiedzialność poznawcza w świetle analizy Naomi S. Baron

AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Artykuł analizuje wpływ generatywnej AI na pojęcie autorstwa, rozróżniając teksty użytkowe od tych o charakterze ekspresyjnym. Autor argumentuje, że podczas gdy automatyzacja instrukcji czy raportów jest korzystna ze względu na standaryzację, zastąpienie człowieka w tekstach osobistych niszczy ich istotę. Wartość wielu wypowiedzi nie wynika z samej formy, lecz z indeksykalności – związku między słowem a konkretną osobą i jej sprawstwem. Przywołując teorię aktów mowy Austina i Searle'a, tekst wskazuje na ryzyko utraty mocy illokucyjnej w komunikacji zapośredniczonej przez AI. Proces pisania zostaje przedstawiony nie tylko jako produkcja treści, ale jako narzędzie samopoznania i kształtowania postaw, którego nie da się zastąpić gotowym rezultatem z maszyny.

The article analyzes the impact of generative AI on the concept of authorship, distinguishing utilitarian texts from those of an expressive nature. The author argues that while the automation of instructions or reports is beneficial due to standardization, replacing humans in personal texts destroys their essence. The value of many utterances does not stem from form alone, but from indexicality—the link between the word and a specific person and their agency. Drawing on the speech act theory of Austin and Searle, the text points to the risk of losing illocutionary force in communication mediated by AI. The writing process is presented not only as content production

but as a tool for self-discovery and the shaping of attitudes, which cannot be replaced by a ready-made result from a machine.

Artykuł podejmuje problem kryzysu autorstwa w dobie generatywnej sztucznej inteligencji, stawiając tezę, że wartość tekstu nie jest immanentną cechą produktu, lecz wynika z pełnionej przez niego funkcji. Autor analizuje kontinuum zależności między treścią a sprawstwem: od tekstów użytkowych, gdzie automatyzacja i standaryzacja są zaletami, po wypowiedzi indeksykalne (np. listy osobiste, świadectwa), w których sens jest nierozerwalnie związany z ludzkim doświadczeniem i odpowiedzialnością. Tekst krytycznie przygląda się iluzji sprawstwa w relacji człowiek-maszyna, ostrzegając przed „pokusą wydajności", która może prowadzić do poznawczego pasożytnictwa i homogenizacji myślenia. Szczególna uwaga zostaje poświęcona sferze akademickiej, gdzie AI obnaża słabość traktowania tekstu jako substytutu procesu myślowego. W konkluzji artykuł postuluje przejście od oceny gotowego artefaktu ku weryfikacji procesu tworzenia, argumentując, że w świecie syntetycznej koherencji to właśnie opór materiału, świadomy wybór i egzystencjalny koszt decyzji pozostają jedynymi autentycznymi wyznacznikami ludzkiego autorstwa.

SŁOWA KLUCZOWE

kryzys autorstwa | sztuczna inteligencja | funkcja tekstu | odpowiedzialność poznawcza | generatywna AI | automatyzacja pisania | moc illokucyjna | homogenizacja poznawcza | P-creativity | H-creativity | intentional fallacy | funkcja autora | epistemiczna odpowiedzialność | prawo Goodharta | strefy wolne od AI

Wartość tekstu zależy od jego funkcji

Kto to napisał? Człowiek, maszyna i kryzys autorstwa w epoce sztucznej inteligencji

Literatura nie jest fakturą: funkcja tekstu, obecność autora i granice automatyzacji

Nie wszystkie teksty wymagają autora w tym samym sensie. Stwierdzenie to wydaje się banalne, dopóki nie spróbujemy konsekwentnie wyprowadzić z niego wniosków dotyczących sztucznej inteligencji.

Paragon, ostrzeżenie meteorologiczne, instrukcja obsługi urządzenia czy raport o temperaturze serwera, a z drugiej strony osobisty list pożegnalny – wszystkie te teksty należą do tej samej kultury pisma, lecz ich wartość opiera się na zupełnie innych relacjach między słowem, światem i człowiekiem. W pierwszych przypadkach oczekujemy przede wszystkim, aby tekst pełnił określoną funkcję: był poprawny, jednoznaczny, aktualny i użyteczny. W przypadku listu pożegnalnego pytanie o to, kto go napisał, nie jest informacją poboczną; stanowi ono istotę sensu wypowiedzi. List wygenerowany przez system na podstawie polecenia „napisz wzruszająco" może być stylistycznie bezbłędny, ale jego znaczenie zmienia się radykalnie, gdy odbiorca dowiaduje się, że osoba podpisana pod tekstem nie sformułowała tych słów ani nie dokonała świadomego wyboru ich użycia. Punktem wyjścia w tej analizie może być taksonomia motywacji do pisania zaproponowana przez Naomi S. Baron. Obejmuje ona m.in. pisanie codzienne, wykonywanie zewnętrznych poleceń, dążenie do korzyści zawodowych, przekazywanie wiedzy, porządkowanie świata, samoodkrycie oraz ekspresję i bunt.

Kategorie te pozwalają dostrzec, że automatyzacja nie jest pojęciem jednolitym. Jeśli celem jest sporządzenie listy produktów lub przekształcenie danych w standardowy raport, istotą zadania pozostaje rezultat. Jeżeli jednak człowiek pisze, aby zrozumieć własny ból, przepracować stratę, ukształtować stanowisko polityczne lub znaleźć język dla swojego doświadczenia, proces tworzenia tekstu staje się częścią celu. Oddanie go maszynie przypomina wtedy wynajęcie kogoś, kto zamiast nas pójdzie na terapię albo odbędzie trening fizyczny. Możemy otrzymać opis rezultatu, ale nie przejdziemy procesu, który miał nas zmienić.

Taka argumentacja nie oznacza, że tekst użytkowy jest gorszy od literackiego lub że literatura posiada metafizyczny monopol na człowieczeństwo. Różnica ma charakter funkcjonalny. Instrukcję obsługi oceniamy przede wszystkim przez pryzmat tego, czy pozwala ona prawidłowo obsłużyć urządzenie. Jeśli maszyna napisze ją jaśniej niż człowiek, trudno wskazać poważny powód, dla którego należałoby preferować wersję ludzką wyłącznie z szacunku dla cierpienia technicznego autora. Można wręcz stwierdzić, że w wielu tekstach użytkowych zbyt silna obecność indywidualnego stylu jest wadą. Procedura bezpieczeństwa nie powinna zachwycać niejednoznaczną metaforą, umowa nie zyskuje na artystycznej tajemniczości, a raport operacyjny nie musi mieć rozpoznawalnego „głosu autora", jeśli utrudnia on jednoznaczny odczyt danych. Im bardziej język pełni funkcję narzędzia, tym łatwiejsza staje się jego automatyzacja. Maszyna może być wręcz uprzywilejowana tam, gdzie wymaga się standardyzacji, powtarzalności i ograniczenia

indywidualnej wariancji. Ta sama cecha, którą krytycy generatywnego pisania nazywają homogenizacją, w innym kontekście staje się zaletą. Ujednolicenie terminologii w instrukcji medycznej, dokumentacji technicznej czy systemie korporacyjnym po prostu zmniejsza ryzyko interpretacyjnego chaosu.

Wartość tekstu zależy od funkcji komunikacyjnej i sprawstwa

Nie istnieje zatem uniwersalne prawo mówiące, że "unikalny ludzki głos" zawsze poprawia tekst. Czasami to właśnie tego głosu chcemy się pozbyć. Sytuacja zmienia się jednak, gdy wartość wypowiedzi posiada komponent indeksykalny: tekst wskazuje na konkretną osobę, jej historię i odpowiedzialność. Zdanie "kocham cię" funkcjonuje inaczej niż informacja "temperatura wynosi 18 stopni".

W drugim przypadku interesuje nas przede wszystkim zgodność zdania ze stanem świata. W pierwszym prawdziwość wypowiedzi zależy także od relacji między słowami a osobą, która je wypowiada. System językowy może prawidłowo skonstruować deklarację miłości, lecz jeżeli odbiorca sądzi, że pochodzi ona od człowieka, który faktycznie jej nie sformułował, zmienia się charakter aktu komunikacyjnego.

Forma pozostaje, ale sprawstwo zostaje przesunięte. Filozofia języka od dawna rozróżniała opis świata od czynności wykonywanych poprzez wypowiedź. John L. Austin pokazywał, że mówiąc, ludzie nie tylko opisują rzeczywistość, lecz również wykonują działania: obiecują, przepraszają, przyrzekają, mianują, ostrzegają. John Searle rozwijał następnie teorię aktów mowy, wskazując warunki, w których wypowiedź posiada określoną moc illokucyjną. Dla naszej analizy konsekwencja jest zasadnicza: niektóre teksty są wartościowe właśnie jako akty osoby, a nie wyłącznie jako układy informacji. Przeprosiny napisane przez system mogą pomóc człowiekowi znaleźć właściwe słowa. Jeśli jednak nadawca nawet ich nie przeczytał i automatycznie nacisnął "wyślij", powstaje pytanie, czy odbiorca otrzymał akt skruchy, czy jedynie jego językową symulację.

Podobny problem dotyczy świadectwa. Zdanie "widziałem to" jest czymś więcej niż opisem sceny; ustanawia relację pomiędzy mówiącym, zdarzeniem i wspólnotą odbiorców. Świadek bierze na siebie odpowiedzialność za prawdziwość relacji. Dlatego autobiografia, zeznanie, reportaż pierwszoosobowy i pamiętnik nie mogą zostać ocenione wyłącznie według kryterium płynności prozy. Jeżeli maszyna stworzy doskonały opis przeżycia wojny, którego nikt z podpisanych autorów

nie doświadczył, problemem nie jest stylistyka zdań — mogą one być przecież poruszające. Problem polega na fałszywej relacji między wypowiedzią a doświadczeniem.

To rozróżnienie pozwala lepiej zrozumieć intuicję Baron dotyczącą "stref wolnych od AI". W materiale źródłowym jako szczególnie chronione wskazane są pamiętniki, dzienniki, korespondencja intymna i krytyczne eseje. Nie trzeba jednak przyjmować tej propozycji jako absolutnego zakazu.

Osoba z niepełnosprawnością może potrzebować technologicznego wsparcia do sformułowania listu osobistego; ktoś piszący w obcym języku może wykorzystać system do korekty bez utraty własnej intencji; autor może poprosić AI o pomoc w uporządkowaniu wspomnień. Istotne pytanie brzmi jednak, czy narzędzie wspiera ekspresję doświadczenia, czy zaczyna je zastępować. Im bardziej tekst ma świadczyć o tym, kim jest mówiący, tym większe znaczenie ma stopień jego rzeczywistego udziału.

Literatura komplikuje ten obraz, ponieważ fikcja z definicji nie wymaga prawdziwości biograficznej. Pisarz może opisywać uczucia, których nigdy nie przeżył, miejsca, których nie widział, albo postać należącą do innej epoki. Nie byłoby rozsądne twierdzić, że Marcel Proust musi osobiście doświadczyć każdej sytuacji swoich bohaterów, aby stworzyć wartościową literaturę. Wyobraźnia jest twórcza właśnie dlatego, że przekracza bezpośrednie doświadczenie jednostki.

Argument "AI nie przeżywa, więc nie może tworzyć literatury" okazuje się zatem słabszy, niż początkowo wygląda. Literatura nigdy nie była prostym stenogramem doznań autora. Co więcej, ogromna część XX-wiecznej teorii literatury celowo osłabiała kult intencji autorskiej. W.K. Wimsatt i Monroe Beardsley określili jako "intentional fallacy" błąd polegający na uzależnianiu wartości i znaczenia dzieła od rekonstruowania zamiaru autora. Roland Barthes w słynnym eseju o "śmierci autora" przesuwał centrum znaczenia od biograficznego źródła ku samemu tekstowi i jego odbiorcy. Michel Foucault proponował z kolei mówić o "funkcji autora": o tym, jak nazwisko porządkuje dyskurs, przypisuje odpowiedzialność i klasyfikuje teksty.

Z perspektywy tych teorii pojawienie się literatury generowanej przez AI może zostać potraktowane niemal jako brutalnie dosłowna realizacja dawnego eksperymentu humanistyki. Oto naprawdę istnieje tekst, który może funkcjonować bez tradycyjnego autora. Jest to jeden z najmocniejszych kontrargumentów wobec obrony ludzkiego pisania i należy go potraktować poważnie. Jeśli wiersz porusza czytelnika, metafora otwiera nową interpretację, a powieść tworzy spójny świat, dlaczego odbiorca miałby odrzucić doświadczenie estetyczne tylko dlatego, że po drugiej stronie nie ma biologicznego twórcy? Czy muzyka traci piękno, gdy nie znamy kompozytora?

Czy anonimowe podanie ludowe jest mniej wartościowe dlatego, że nie potrafimy wskazać jednej intencjonalnej jednostki odpowiedzialnej za każdą jego wersję? Kultura zna przecież teksty zbiorowe, anonimowe, redagowane przez pokolenia, a nawet powstające częściowo przez przypadek. Współczesne eksperymenty pokazują, jak poważne jest to wyzwanie.

Wartość tekstu zależy od procesu i oporu

Badanie opublikowane w "Scientific Reports" w 2024 roku wykazało, że osoby niebędące ekspertami nie tylko nie potrafiły wiarygodnie odróżnić poezji generowanej przez AI od utworów znanych poetów anglojęzycznych, lecz w wielu wymiarach oceniały teksty maszynowe wyżej. Uczestnicy częściej uznawali wiersze AI za ludzkie niż rzeczywiste utwory ludzi. Jednocześnie informacja o tym, że tekst został wygenerowany przez sztuczną inteligencję, obniżała jego ocenę. Mamy więc do czynienia z niezwykle interesującym eksperymentem naturalnym: wartość tekstu zależy częściowo od samej treści, a częściowo od opowieści o jego pochodzeniu. Nie jest to jednak dowód na irracjonalność odbiorców.

Może to wskazywać na to, że sztukę oceniamy jednocześnie na kilku poziomach. Pierwszy dotyczy właściwości artefaktu: rytmu, obrazu, narracji i emocjonalnej czytelności. Drugi odnosi się do procesu powstawania: trudności, intencji, doświadczenia oraz ryzyka. Trzeci dotyczy relacji społecznej – tego, czy dzieło pozwala nam spotkać się z cudzym sposobem widzenia świata. AI może skutecznie konkurować na poziomie pierwszym, a mimo to pozostawiać otwarte pytania na dwóch pozostałych. Co więcej, wyniki badań sugerują, że generatywna AI może przewyższać człowieka właśnie w tych cechach, które masowy odbiorca utożsamia z "dobrym tekstem": klarowności, dostępności oraz bezpośredniego komunikowania emocji i tematu. Autorzy badania zauważają, że skomplikowane utwory ludzkich poetów wymagają większego wysiłku interpretacyjnego, podczas gdy tekst AI łatwiej przekazuje obraz lub nastrój. W tym miejscu powraca "pokusa wydajności", lecz tym razem po stronie czytelnika. Nie tylko autor może preferować szybkie pisanie; odbiorca może preferować szybką interpretację.

Prowadzi to do przewrotnego pytania: czy sztuczna inteligencja będzie produkować literaturę gorszą, czy przede wszystkim taką, która jest łatwiej konsumowalna? Jeśli modele uczą się statystycznych preferencji i są dostrajane tak, by generować treści pomocne, jasne i satysfakcjonujące, mogą naturalnie dążyć ku językowi pozbawionemu części twórczego oporu. Tymczasem wielka literatura często działa odwrotnie. Kafka nie jest wartościowy dlatego, że optymalnie minimalizuje wysiłek odbiorcy. Beckett nie projektuje doświadczenia bez tarcia. Poezja Celana nie aspiruje do natychmiastowej jednoznaczności. Nieprzejrzystość bywa narzędziem

artystycznym, ponieważ zmusza czytelnika do pracy interpretacyjnej. Tutaj pojawia się możliwość odwrócenia kryteriów.

W świecie przed generatywną AI gładkość tekstu mogła być świadectwem kunsztu. W erze masowej generacji to właśnie pęknięcie, ryzyko, osobliwość i opór mogą nabrać większej wartości jako sygnały nieprzewidywalnego autorstwa. Nie chodzi o romantyczny kult błędu – źle napisane zdanie nie staje się dziełem sztuki tylko dlatego, że maszyna napisałaby je lepiej. Chodzi o różnicę między niedoskonałością przypadkową a świadomym przekroczeniem normy. Wielcy pisarze często łamali reguły języka, gdyż te nie wystarczały do wyrażenia ich zamysłu. Model uczący się na dorobku kultury może symulować takie naruszenia; pytanie brzmi jednak, czy potrafi ustanowić nową normę kulturową, a nie jedynie statystycznie odtworzyć ślady wcześniejszych buntów.

W literaturze szczególną wartość posiada styl. Nie jest on ozdobną warstwą nakładaną na treść, lecz sposobem selekcji świata. Proustowskie zdanie, lapidarność Hemingwaya, fragmentaryczność modernistów czy składniowe eksperymenty Joyce'a nie są wymiennymi dekoracjami – one określają, co i jak może zostać zauważone. Jeśli AI potrafi naśladować styl, dowodzi ogromnej kompetencji w modelowaniu formy, ale nie rozstrzyga, czy potrafi stworzyć styl historycznie nowy, wynikający z konfliktu między dotychczasowym językiem a doświadczeniem, dla którego istniejący język okazuje się niewystarczający.

Tu dochodzimy do kategorii autentyczności. Jest ona pojęciem niebezpiecznym, gdyż łatwo uczynić z niej mglistą pochwałę wszystkiego, co ludzkie. Człowiek potrafi być bowiem skrajnie nieautentyczny, wtórny i schematyczny. Maszyna natomiast może stworzyć tekst zaskakujący, oryginalny w odbiorze i estetycznie skuteczny.

Autentyczność jako zgodność procesu z deklaracją

Autentyczność nie powinna więc oznaczać po prostu biologicznego pochodzenia. Bardziej użyteczne jest rozumienie jej jako zgodności między deklarowanym źródłem wypowiedzi a rzeczywistym procesem jej powstawania. Tekst staje się problematycznie nieautentyczny nie wtedy, gdy pomagała w nim maszyna, lecz gdy przedstawia się go odbiorcy jako bezpośredni wyraz czyjegoś doświadczenia, choć faktyczny związek osoby z wypowiedzią jest znacznie słabszy. Badania nad narracjami generowanymi przez AI pokazują zresztą, że sama etykieta autorstwa może wpływać na odbiór niezależnie od rzeczywistego pochodzenia tekstu. W eksperymentach z 2025 roku deklarowane ludzkie autorstwo potrafiło podnosić określone oceny lub preferencje, podczas gdy faktyczna różnica między tekstem człowieka a systemu nie zawsze przekładała się na równie silne różnice w zaangażowaniu czy przyjemności lektury.

Oznacza to, że "autor" działa także jako informacja kontekstowa. Czytelnik nie czyta wyłącznie zdań – czyta je jako czyjeś zdania. Możemy zatem zaproponować kontinuum zależności wartości tekstu od autora. Na jednym krańcu znajdują się teksty czysto funkcjonalne, w których autorstwo może być niemal całkowicie obojętne, o ile zapewniono prawdziwość i odpowiedzialność instytucjonalną. W środku plasują się analizy, artykuły naukowe, komentarze prawne czy publicystyka, gdzie liczy się zarówno rezultat, jak i możliwość wskazania osoby odpowiedzialnej za selekcję argumentów. Na drugim krańcu umieścimy formy, w których tożsamość wypowiadającego jest integralna dla sensu: zeznanie, pamiętnik, osobiste przeprosiny, list intymny, autobiografia. Literatura pozostaje paradoksalnie pomiędzy nimi. Może być interpretowana autonomicznie wobec autora, a jednocześnie kultura uporczywie chce wiedzieć, czy stoi za nią drugi człowiek. To właśnie ta uporczywość jest ważna. Człowiek nie tylko konsumuje komunikaty – poszukuje innych umysłów.

Czytanie jest jednym z mechanizmów przekraczania granic własnego doświadczenia: przez kilka godzin można patrzeć na świat poprzez język innej osoby. Jeśli okaże się, że "inna osoba" nie istniała jako podmiot danego tekstu, część odbiorców nadal zachowa pełne doświadczenie estetyczne. Inni poczują jednak, że zmienił się przedmiot doświadczenia. Nie oznacza to, że tekst nagle utracił wszystkie właściwości; oznacza raczej, że literatura była dla nich nie tylko wzorem znaków, lecz formą spotkania.

Słynna "śmierć autora" może w epoce AI wymagać zaskakującej korekty. Teoria literatury chciała uwolnić tekst od tyranii biograficznej intencji, ale niekoniecznie przewidywała kulturę, w której tekst może powstać bez żadnego człowieka wykonującego centralny akt kompozycji. Różnica między "nie muszę znać intencji autora, aby interpretować dzieło" a "autor w tradycyjnym sensie nie istniał" jest większa, niż może się wydawać. Pierwsze twierdzenie ogranicza władzę autora nad znaczeniem; drugie podważa samą społeczną relację, którą część odbiorców uważała za źródło wartości sztuki.

Nie musimy jeszcze rozstrzygać tego sporu. Ważniejsze jest zauważenie, że generatywna AI niszczy możliwość posługiwania się jednym kryterium dla całej kultury pisma. Automatycznie wygenerowana faktura może być doskonała właśnie dlatego, że nikt nie oczekuje od niej biografii. Automatycznie wygenerowane przeprosiny mogą być moralnie puste właśnie dlatego, że biografia i odpowiedzialność są ich treścią. Automatycznie wygenerowany wiersz może zaś znajdować się w najbardziej interesującym miejscu pomiędzy tymi biegunami: być estetycznie skuteczny, interpretacyjnie płodny i jednocześnie pozbawiony tego rodzaju ludzkiego źródła, które przez stulecia przyjmowaliśmy jako oczywiste. Czas wejść dokładnie w tę strefę pośrednią.

Kreatywność jako cecha produktu a sprawstwo podmiotu

Pytanie nie będzie już brzmiało, czy maszyna potrafi stworzyć coś pięknego – empirycznie wiemy, że odbiorcy uznają jej wytwory za wartościowe. Trudniejszy problem dotyczy tego, czy kreatywność jest właściwością produktu, procesu, twórcy, czy może relacji między nimi. Koncepcje P-creativity i H-creativity Margaret Boden, model 4C oraz współczesne eksperymenty generatywne pozwalają zbadać, czy maszyna może być nie tylko producentem udanego tekstu, lecz także uczestnikiem procesu, który kultura nazywa twórczością. Kreatywność bez przeżycia: Boden, Big-C i przypadkowy opad symboli. Najtrudniejszy argument przeciwko prostemu przeciwstawieniu człowieka i maszyny pojawia się właśnie wtedy, gdy sztuczna inteligencja robi coś, co wygląda na twórcze.

Gdy system poprawia ortografię, możemy spokojnie nazwać go narzędziem. Gdy streszcza dokument, mówimy o automatyzacji. Kiedy jednak tworzy metaforę, melodię, obraz albo zakończenie opowiadania, język zaczyna się chwiać. Czy maszyna "wymyśliła" rozwiązanie? Czy była "kreatywna"? A może jedynie wytworzyła konfigurację, której wcześniej nie widzieliśmy?

Problem jest poważniejszy, niż sugeruje popularne pytanie o to, czy AI może "zostać artystą", ponieważ samo pojęcie kreatywności od dawna nie posiada jednej, powszechnie zaakceptowanej definicji. W psychologii najczęściej łączy się dwa kryteria: nowość oraz wartość lub użyteczność. Margaret Boden ujmuje twórczość właśnie poprzez możliwość generowania idei nowych, zaskakujących i wartościowych, rozróżniając przy tym nowość psychologiczną od historycznej. Jeżeli przyjmiemy wyłącznie kryterium produktu, część wytworów systemów generatywnych niewątpliwie kwalifikuje się jako twórcza. Jeśli jednak do definicji dodamy intencję, samowiedzę, motywację, biografię i rozwój podmiotu, odpowiedź przestaje być oczywista.

Boden wprowadza rozróżnienie między P-creativity a H-creativity. Pierwsza oznacza ideę nową dla konkretnego umysłu, nawet jeśli ludzkość znała ją wcześniej. Dziecko samodzielnie odkrywające zasadę matematyczną może być kreatywne w ten sposób, choć matematyk opisał ją trzy stulecia wcześniej. H-creativity oznacza natomiast nowość historyczną: ideę, której – według dostępnej wiedzy – wcześniej w kulturze nie było. Rozróżnienie to jest wyjątkowo przydatne w epoce AI, ponieważ odsłania ukryty problem. H-kreatywność można w pewnym stopniu oceniać z zewnątrz, porównując rezultat z historią danej dziedziny. P-kreatywność odsyła natomiast do historii konkretnego podmiotu. Aby uznać, że coś było psychologicznie nowe dla autora, musimy przyjąć, że posiada on określony stan wiedzy, doświadczenia i własną trajektorię poznawczą.

W przypadku modelu językowego takie przeniesienie pojęcia staje się metaforyczne. Możemy mówić, że wynik jest nowy względem innych wyników systemu, ale nie wynika z tego, że maszyna doświadczyła własnego odkrycia jako odkrycia. Podobna ostrożność jest potrzebna przy analizie modelu 4C Jamesa Kaufmana i Ronalda Beghetto.

Autorzy rozbudowali tradycyjne przeciwstawienie codziennej "little-c" i wybitnej "Big-C" kreatywności o "mini-c", charakterystyczną dla osobistego procesu uczenia się i konstruowania nowych znaczeń, oraz "Pro-c", czyli twórczość rozwiniętą poprzez długoletnią praktykę do poziomu zawodowej ekspertyzy. Model ten pozwala ocenić sztuczną inteligencję, jednak wymaga korekty metodologicznej: 4C powstało jako teoria rozwoju ludzkiej kreatywności, a nie skala przeznaczona do klasyfikowania systemów generatywnych. Szczególnie mini-c zakłada zmianę osobistego rozumienia, a Pro-c lata nabywania wiedzy i funkcjonowania w społeczności zawodowej. Twierdzenie, że model "znajduje się na poziomie Pro-c", może więc oznaczać podobieństwo produktu do wytworów profesjonalisty, ale nie tożsamość procesów prowadzących do tej jakości.

Uwidacznia się tu zasadniczy konflikt między kreatywnością produktu a kreatywnością podmiotu. Wyobraźmy sobie melodię, która nigdy wcześniej nie została skomponowana, jest estetycznie wartościowa i zostaje wysoko oceniona przez kompetentnych słuchaczy. Jeżeli definicja twórczości dotyczy wyłącznie artefaktu, sprawa wydaje się rozstrzygnięta: powstał twórczy produkt. Jeśli jednak pytamy, czy jego generator jest twórcą, potrzebujemy dodatkowych kryteriów. Czy musiał posiadać cel estetyczny? Czy powinien rozpoznawać, co jest zaskakujące? Czy musi mieć możliwość niezadowolenia z pierwszej wersji?

Czy powinien wiedzieć, że przekroczył normę?

Formalna kompetencja maszyny demistyfikuje pojęcie stylu

Czy twórczość wymaga tego, co filozofia działania nazywa sprawstwem, czyli zdolności do przypisania sobie działania jako własnego? Generatywna AI rozdziela te pytania, które w przypadku człowieka zwykle zlewały się w jedną figurę artysty. Znakomitym przykładem wcześniejszego eksperymentu tego rodzaju są prace Davida Cope'a.

Jego system Experiments in Musical Intelligence, rozwijany od lat osiemdziesiątych, analizował utwory konkretnych kompozytorów, identyfikował charakterystyczne wzorce, a następnie rozkładał materiał na elementy i rekombinował je w nowe kompozycje zachowujące cechy danego stylu. Uniwersytet Kalifornijski w Santa Cruz opisywał EMI jako program tworzący oryginalne utwory w

stylach takich twórców jak Bach, Mozart czy Strawiński. Rezultaty były na tyle przekonujące, że wywołały spór nie tylko o technologię, lecz o samą naturę muzycznej kreatywności. Douglas Hofstadter przyznawał, że doświadczenie muzyki wytworzonej przez stosunkowo ograniczony mechanizm zmusiło go do ponownego zastanowienia się nad tym, co wcześniej uważał za głęboko ludzką ekspresję. To szczególnie niewygodna lekcja dla humanizmu opartego na intuicji, że skomplikowany rezultat musi świadczyć o równie skomplikowanym wnętrzu twórcy. Być może część tego, co nazywamy stylem, jest bardziej formalizowalna, niż chcielibyśmy przyznać. Jeśli wystarczająco precyzyjnie zidentyfikujemy regularności harmoniczne, rytmiczne, składniowe czy metaforyczne, możemy produkować artefakty rozpoznawane przez odbiorców jako przynależne do określonej tradycji.

Nie oznacza to, że Bach "był algorytmem". Oznacza raczej, że część zjawisk, które słuchacz interpretował jako ślad indywidualnego geniuszu, może posiadać strukturę możliwą do modelowania. Maszyna staje się narzędziem demistyfikującym: niekoniecznie wyjaśnia całą kreatywność, ale pokazuje, że pewne jej powierzchnie można odtworzyć bez rekonstruowania całej psychologii autora. Podobnych doświadczeń dostarczył w poezji Deep-speare. Model opisany w 2018 roku skonstruowano tak, aby uwzględniał język, metrum i rym sonetu. W zakresie elementów formalnych wygenerowane utwory osiągały wyniki na tyle dobre, że trudno było je odróżnić od tekstów ludzkich. Eksperci wskazywali jednak na wyraźniejsze słabości w obszarze czytelności oraz emocjonalności treści. Nie jest to drobiazg.

System może sprawnie opanować ograniczenia formy, a jednocześnie dowodzić, że poezja nie wyczerpuje się w samym spełnianiu metrum i schematu rymów. Tym samym eksperyment nie rozstrzyga pytania o kreatywność maszyny, lecz rozbija je na komponenty: formalną kompetencję, językową nowość, emocjonalny efekt, spójność, znaczenie i historyczną doniosłość. Podobne pytania pojawiły się na rynku sztuki w 2018 roku wraz z "Portrait of Edmond de Belamy". Praca przygotowana przez francuski kolektyw Obvious z wykorzystaniem sieci GAN została sprzedana w Christie's za 432 500 dolarów, wielokrotnie przewyższając estymację aukcyjną. Najciekawszy w tym wydarzeniu nie był jednak sam wynik finansowy.

Kreatywność jako systemowa relacja człowieka i narzędzia

Już wtedy jeden z członków kolektywu uchwycił paradoks autorstwa: jeśli artystą jest podmiot fizycznie generujący obraz, należałoby wskazać maszynę; jeśli zaś jest nim ten, kto posiada wizję i chce przekazać komunikat, rolę tę zachowują ludzie. Słynny portret, zamiast udowodnić, że „AI

została artystą", ujawnił zatem trudność w ustaleniu jednostki analizy. Dzieło okazało się rezultatem relacji między wcześniejszymi obrazami, architekturą algorytmiczną, twórcami oprogramowania, kolektywem dokonującym selekcji oraz rynkiem, który nadał efektowi końcowemu status sztuki.

Prowadzi to ku teorii kreatywności jako zjawiska systemowego. Mihaly Csikszentmihalyi już przed erą LLM argumentował, że kreatywność nie powstaje wyłącznie „w głowie" jednostki. Wymaga ona relacji między twórcą, domeną kultury a polem ekspertów, którzy rozpoznają innowację jako wartościową. Z tej perspektywy nawet ludzkie Big-C nie jest cechą czysto neuronalną. Mozart nie stałby się „Mozartem" bez systemu tonalnego, tradycji muzycznej, wykonawców, instytucji, odbiorców i późniejszej historii recepcji. Generatywna AI wzmacnia więc stare pytanie: skoro kreatywność jest rozproszona między twórcę, kulturę i odbiorców, dlaczego część tej sieci nie mogłaby być niebiologiczna? Odpowiedź nie jest oczywista, gdyż można argumentować, że system AI wcale nie musi być samodzielnym artystą, aby stać się częścią układu twórczego.

Aparat fotograficzny nie ma intencji, a jednak zmienił historię sztuki. Syntezator nie odczuwa muzyki, lecz rozszerzył przestrzeń możliwych brzmień. Photoshop nie posiada estetycznych pragnień, a mimo to stał się medium twórczym. W tym ujęciu modele generatywne można postrzegać jako wyjątkowo aktywne narzędzia: zamiast jedynie wykonywać polecenia człowieka, generują przestrzeń możliwości, której twórca nie byłby w stanie samodzielnie tak szybko eksplorować. Nowość nie musi więc należeć wyłącznie do systemu albo człowieka – może powstawać pomiędzy nimi.

Współczesne badania nad human-AI co-creation obnażają jednak pewną ambiwalencję. W eksperymentach dotyczących pisania poezji z 2024 roku McGuire, De Cremer i Van de Cruys stwierdzili, że sama obecność zaawansowanego narzędzia AI nie gwarantowała wzrostu ludzkiej kreatywności. Uczestnicy piszący samodzielnie uzyskiwali bardziej kreatywne rezultaty niż osoby rozpoczynające od gotowego wiersza wygenerowanego przez AI, który następnie edytowały; autorzy podkreślali tu znaczenie poczucia współtworzenia i własnej skuteczności.

Jest to istotne, gdyż pokazuje, że ogromna moc generatywna narzędzia może paradoksalnie zmniejszyć wkład użytkownika. Gdy system podaje produkt zbyt wcześnie, człowiek zostaje zakotwiczony w cudzym rozwiązaniu. Inne badania przynoszą bardziej optymistyczne wnioski. Eksperyment z 2026 roku opublikowany w „Scientific Reports", obejmujący 92 mężczyzn uczących się języka angielskiego, wykazał poprawę wyników kreatywnego pisania w grupach korzystających ze wsparcia AI, szczególnie przy rozbudowanym sprzężeniu zwrotnym. Uczestnicy deklarowali, że sugestie systemu otwierały przed nimi nowe możliwości narracyjne.

Badanie to ma jednak wąski zakres populacyjny i dydaktyczny, więc nie pozwala twierdzić, że AI uniwersalnie „zwiększa kreatywność". Wskazuje raczej na coś subtelniejszego: narzędzie może pełnić rolę zewnętrznego generatora alternatyw, zwłaszcza tam, gdzie brak płynności językowej blokuje dostęp do procesu twórczego. Inne eksperymenty ujawniają konflikt między wynikiem a psychologią twórcy. Seria czterech badań z 2025 roku, obejmująca 3562 uczestników, wykazała, że współpraca z generatywną AI zwiększa bieżącą efektywność w zadaniach, lecz poprawa ta nie przenosiła się na późniejsze prace wykonywane samodzielnie; jednocześnie obserwowano osłabienie motywacji wewnętrznej. To niemal laboratoryjna ilustracja „pokusy wydajności" Baron. Produkt staje się lepszy, ale niekoniecznie rozwija się osoba, która go podpisuje. Jeśli kreatywność rozumiemy jako jakość artefaktu – AI pomaga. Jeśli jednak postrzegamy ją jako trwałą zdolność i motywację jednostki, bilans wygląda inaczej.

Z kolei badania porównujące LLM z ludźmi w testach myślenia dywergencyjnego dostarczają wyników niewygodnych dla osób chcących zachować prostą granicę antropologiczną. Już praca z 2024 roku wykazała bardzo wysokie wyniki modeli w zadaniach Alternate Uses Task, a nowsze analizy porównujące wiele LLM z ogromnymi próbami ludzkimi potwierdzają, że systemy potrafią osiągać wysoką semantyczną dywergencję i generować nietypowe skojarzenia. Trzeba jednak precyzyjnie określić, co dokładnie zostało zmierzone.

Kreatywność transformacyjna i egzystencjalny koszt wyboru

Test dywergencyjnego myślenia mierzy istotny komponent kreatywności, lecz nie oddaje pełni twórczego życia. Nie uwzględnia dwudziestu lat pracy nad problemem, umiejętności rozpoznania pomysłu wartego rozwoju, społecznego ryzyka związanego ze sprzeciwieniem się własnej dyscyplinie ani decyzji o porzuceniu eleganckiej teorii w obliczu niewygodnego faktu. Dobrym przykładem tej różnicy jest nauka.

Badanie z 2025 roku analizowało zdolność generatywnej AI do przeprowadzenia procesu odkrycia naukowego w warunkach symulujących badania nad genetyką molekularną. Autorzy zauważyli, że systemy radziły sobie z odkryciami inkrementalnymi, opierając się na istniejących reprezentacjach wiedzy, lecz nie odtworzyły kluczowej ludzkiej zdolności do dostrzeżenia anomalii i stworzenia nowej hipotezy „od zera". Nie dowodzi to metafizycznej niemożliwości maszynowego odkrycia, lecz pokazuje, że płynna generacja wariantów a prawdziwie transformacyjne odkrycie to odrębne kompetencje. Margaret Boden rozróżniała zresztą trzy mechanizmy kreatywności: kombinacyjną, eksploracyjną i transformacyjną. Pierwsza łączy istniejące elementy w nieoczekiwany sposób; druga

eksploruje przestrzeń możliwości wyznaczoną przez określone reguły; trzecia zaś zmienia same reguły tej przestrzeni, czyniąc możliwym to, co wcześniej było niewyobrażalne.

Taka typologia pozwala uniknąć jałowego sporu o to, czy rekombinacja „jest naprawdę kreatywna". Znaczna część ludzkiej twórczości również opiera się na rekombinacji. Shakespeare nie wynalazł miłości ani zazdrości i nie stworzył wszystkich fabuł, którymi się posługiwał. Bach nie powołał systemu tonalnego z niczego, a Einstein nie wymyślił matematyki, na której budował fizykę. Twórczość nie oznacza kreacji ex nihilo; pytanie brzmi raczej, kiedy eksploracja istniejącej przestrzeni staje się jej transformacją.

W kreatywności transformacyjnej ludzka biografia może mieć szczególne znaczenie. Człowiek nie tylko przelicza warianty – on doświadcza nieadekwatności dotychczasowych kategorii. Artysta czuje, że odziedziczona forma nie pozwala mu wyrazić tego, co istotne; naukowiec dostrzega anomalię, której teoria nie wyjaśnia; myśliciel odkrywa problem moralny w tym, co jego epoka uznawała za oczywistość. Transformacja rodzi się więc z oporu świata wobec naszych pojęć. Jest to inny rodzaj nowości niż losowe odejście od średniej, co przywołuje określenie Geoffreya Jeffersona o „przypadkowym układzie symboli".

W materiale Baron zwrot ten podkreśla różnicę między sonetem zrodzonym z przeżycia a sonetem wytworzonym przez system. Współczesnych modeli nie należy jednak utożsamiać z generatorami czystego przypadku; ich wyniki są ustrukturyzowane przez wyuczone regularności, kontekst i optymalizację. „Przypadkowość" w próbkowaniu nie wyjaśnia kompetencji systemu. Aktualna wersja problemu Jeffersona jest subtelniejsza: czy nowość wygenerowana z modelu kultury jest tym samym rodzajem nowości, co ta powstająca z konfrontacji żyjącego podmiotu z własnym światem? To pytanie przesuwa punkt ciężkości z psychologii kreatywności ku fenomenologii.

Człowiek tworzy jako istota zanurzona w czasie. Ma dzieciństwo, ciało, ograniczoną przyszłość i lęk przed śmiercią; ma relacje, wstyd, ambicję, zmęczenie oraz świadomość, że pewnych decyzji nie można cofnąć. Ta skończoność sprawia, że wybór twórczy niesie ze sobą koszt. Pisarz poświęcający dekadę jednej książce rezygnuje z innych możliwych tekstów i z części własnego życia.

System może wygenerować tysiąc wariantów i nie ponosi egzystencjalnego kosztu odrzucenia dziewięciuset dziewięćdziesięciu dziewięciu. Nie oznacza to, że wybrany rezultat jest estetycznie bezwartościowy, lecz że słowo „wybór" opisuje w obu przypadkach zupełnie inne zjawisko. W tym miejscu warto przywołać marksowskie pojęcie alienacji, zachowując jednak historyczną ostrożność. Marks analizował sytuację utraty kontroli pracownika nad produktem i własną sprawczością.

W pisaniu wspomaganym przez AI może pojawić się analogiczny fenomen psychologiczny, nawet przy innych warunkach technologicznych. Przykład Jennifer Lepp korzystającej z Sudowrite

pokazuje, że wzrost produktywności współistniał z poczuciem utraty intymnej więzi z postaciami i tekstem. Nie jest to dowód uniwersalnej alienacji, lecz istotne studium przypadku: sukces mierzony liczbą napisanych stron może współistnieć z porażką mierzoną poczuciem autorstwa.

Kreatywność jako proces w układzie człowiek-maszyna

Można wyobrazić sobie przeciwną sytuację. Pisarz wykorzystuje model nie po to, by otrzymywać gotowe sceny, lecz by prowokować własną wyobraźnię. Prosi o najbardziej oczywiste rozwiązania tylko po to, aby je świadomie odrzucić; każe systemowi krytykować konstrukcję bohatera. Generuje dziesięć możliwych motywacji i żadnej nie przyjmuje wprost, lecz jedenastą wymyśla właśnie dzięki oporowi wobec otrzymanych propozycji.

W takim układzie sztuczna inteligencja pełni rolę partnera do burzy mózgów, redaktora lub przypadkowej rozmowy, która staje się impulsem do odkrycia. Produkt nie jest wtedy ani „ludzki" w dziewiętnastowiecznym sensie samotnego geniusza, ani „maszynowy" – jest rezultatem systemu twórczego. Dlatego na obecnym etapie bardziej precyzyjne od pytania o kreatywność AI jest pytanie o lokalizację tej cechy. Czy znajduje się ona w wynikach? W procesie generowania? W osobie wybierającej? W społeczności uznającej dzieło? A może w całym układzie człowiek-model-dane-kultura?

W zależności od przyjętej teorii można udzielić różnych, a zarazem spójnych odpowiedzi. Pod kątem produktu systemy generatywne mogą osiągać wyniki spełniające kryteria nowości i wartości. Psychologicznie nie ma jednak podstaw, by bez dodatkowych założeń przypisywać im ludzką P-kreatywność. Historycznie mogą uczestniczyć w powstawaniu rezultatów H-kreatywnych, choć ustalenie, komu przypisać transformację, będzie coraz trudniejsze. Najrozsądniejsza granica nie przebiega więc między „kreatywnym człowiekiem" a „niekreatywną maszyną", lecz między różnymi znaczeniami twórczości. AI może być wyjątkowo skuteczna w generowaniu wariantów, odległych skojarzeń i przełamywaniu blokady pierwszej wersji. Człowiek zachowuje kluczową rolę tam, gdzie trzeba rozstrzygnąć, dlaczego dana możliwość jest warta realizacji, jakie ma znaczenie wobec przeżywanego świata i czy rezultat zasługuje na odpowiedzialne przedstawienie innym ludziom.

Nie jest to trwały traktat graniczny z technologią, lecz opis aktualnego podziału funkcji, który może ulec zmianie. Kolejny etap analizy powinien odejść od abstrakcyjnego pojedynku „człowiek kontra AI". W realnym świecie coraz rzadziej spotykamy czysto ludzki tekst lub całkowicie autonomiczną produkcję maszyny. Powstają formy pośrednie: sugestia słowa, automatyczne dokończenie zdania, generowanie planu czy wspólne redagowanie. Z każdym takim poziomem zmienia się rozkład sprawstwa.

Pytanie brzmi zatem nie o to, czy maszyna jest twórcza, lecz kto prowadzi w procesie współtworzenia. Czy AI jest lokajem Jeevesem, który dyskretnie poprawia kołnierzyk autora, centaurem rozszerzającym jego możliwości, czy też poznawczym pasożytem, który stopniowo przejmuje decyzje, pozostawiając człowiekowi tylko podpis? Jeeves, centaur czy pasożyt poznawczy?

Najważniejsze pytanie o współpracę człowieka ze sztuczną inteligencją nie brzmi: „czy korzystałeś z AI?".

Przejście od pasywnej korekty do proaktywnej predykcji

Taka informacja jest zbyt uboga, by powiedzieć cokolwiek istotnego o rzeczywistym autorstwie. Dwie osoby mogą odpowiedzieć twierdząco, choć w pierwszym przypadku system poprawił jedynie kilka literówek, a w drugim zaprojektował tezę, argumentację, przykłady, styl i wnioski całego tekstu. Z perspektywy technicznej obie skorzystały ze sztucznej inteligencji; z perspektywy poznawczej, etycznej i autorskiej wykonały jednak zupełnie inne czynności. Właśnie dlatego kwestię współpracy należy opisywać nie przez samą obecność narzędzia, lecz przez rozkład inicjatywy, kontroli i odpowiedzialności.

Naomi S. Baron proponuje w tym kontekście sugestywną figurę Jeevesa – idealnego lokaja znanego z prozy P.G. Wodehouse'a. Współczesny asystent pisania ma być właśnie takim dyskretnym sługą: poprawia kołnierzyk zdania, podpowiada bardziej eleganckie słowo, usuwa niezręczność i zazwyczaj nie rości sobie prawa do autorstwa. Problem zaczyna się wtedy, gdy Jeeves przestaje jedynie służyć panu domu, a zaczyna decydować, dokąd ten ma pójść, co powiedzieć i jak rozumieć własne zamiary. Ewolucja ta jest przejściem od pasywnej korekty do proaktywnej predykcji: od spellcheckera reagującego na gotowy tekst, po systemy takie jak Smart Compose, Grammarly, Microsoft Editor czy Wordtune, które proponują ciąg dalszy, styl i kierunek wypowiedzi. Różnica ta jest głębsza niż zwykły rozwój funkcjonalności interfejsu.

Klasyczny korektor działa po decyzji autora – najpierw człowiek wybiera słowo, potem system zgłasza błąd. Asystent predykcyjny może natomiast zadziałać przed pełnym sformułowaniem myśli. Gdy autor zaczyna pisać „Chciałbym podkreślić…", system już proponuje to, co prawdopodobnie powinno pojawić się dalej. W tym momencie technologia przestaje wyłącznie kontrolować produkt i zaczyna uczestniczyć w konstruowaniu przestrzeni możliwych następstw. Pisarz nie widzi już nieskończonego horyzontu własnego języka, lecz kilka gotowych ścieżek. Może oczywiście każdą z nich odrzucić, jednak sama obecność podpowiedzi zmienia poznawcze środowisko decyzji.

Pojęcie mixed-initiative interaction, rozwijane w badaniach nad interakcją człowiek-komputer już pod koniec XX wieku, okazuje się dziś wyjątkowo aktualne. Eric Horvitz przeciwstawiał proste marzenie o pełnej automatyzacji równie naiwnemu pragnieniu całkowitej kontroli manualnej. Interakcja mieszanej inicjatywy miała umożliwiać elastyczne przekazywanie sterów temu uczestnikowi systemu, który w danym momencie lepiej radzi sobie z zadaniem, przy zachowaniu możliwości ingerencji drugiej strony. Rozwiązanie to brzmi zdroworozsądkowo, lecz w procesie pisania rodzi istotną trudność: jak ustalić moment, w którym asystent powinien podpowiedzieć, a kiedy milczeć? Milczenie narzędzia może być bowiem funkcją równie ważną, co jego inteligencja. Jeśli system dostarcza odpowiedź natychmiast, użytkownik nie musi najpierw doświadczyć własnej niewiedzy. Jeżeli proponuje metaforę, zanim autor sam spróbuje ją znaleźć, likwiduje etap poszukiwania. Jeżeli buduje kontrargument, zanim człowiek zastanowi się nad perspektywą przeciwnika, przejmuje fragment procesu zmiany punktu widzenia. W tradycyjnym projektowaniu oprogramowania szybkość odpowiedzi jest niemal zawsze zaletą. W systemie poznawczym bywa wadą, ponieważ krótkie opóźnienie, frustracja czy konieczność samodzielnego sformułowania stanowiska stanowią integralną część procesu myślenia.

Realna kontrola nad AI wymaga walki z iluzją sprawstwa

Ben Shneiderman buduje koncepcję Human-Centered AI w opozycji do fałszywej alternatywy między autonomią maszyny a autonomią człowieka. Proponuje projektowanie systemów, które będą jednocześnie wysoce zautomatyzowane i pozostaną pod znaczącą ludzką kontrolą, wspierając tym samym samoskuteczność, odpowiedzialność, kreatywność oraz zrozumienie działania narzędzia. Dobre systemy powinny być przewidywalne, sterowalne i zaprojektowane tak, aby człowiek zachowywał realną, a nie ceremonialną kontrolę. W kontekście pisania oznacza to coś więcej niż obecność przycisku "cofnij". Chodzi o zachowanie ludzkiej zdolności do definiowania celu, kryteriów i granic zadania. Materiał źródłowy upraszcza ten problem do kontrastu między modelami „human in the loop” a „AI in the loop”: w pierwszym człowiek sprawuje nadzór, w drugim algorytm wyznacza kierunek, a człowiek staje się jedynie redaktorem jego propozycji.

To rozróżnienie jest użyteczne, choć rzeczywiste systemy oferują znacznie więcej konfiguracji. Możemy mieć człowieka inicjującego i maszynę wykonującą; maszynę proponującą i człowieka selekcjonującego; system generujący warianty i człowieka ustanawiającego kryteria oceny; lub człowieka piszącego pierwszą wersję, gdzie AI pełni funkcję krytyka. Możliwa jest także sytuacja, w której jeden model generuje treść, drugi ją ocenia, a człowiek jedynie zatwierdza wynik. Nie wystarczy więc stwierdzić, że człowiek „pozostaje w pętli” – trzeba zapytać, co właściwie w tej pętli robi. Ma to kluczowe znaczenie, gdyż formalna obecność człowieka może maskować faktyczną

utratę sprawstwa. Operator systemu może naciskać przycisk „zatwierdź”, lecz jeśli brakuje mu czasu, wiedzy lub swobody niezbędnej do rzeczywistej odmowy, jego kontrola staje się iluzoryczna. W badaniach nad automatyzacją zjawisko to znane jest jako automation bias: skłonność do nadmiernego polegania na rekomendacjach systemu nawet wtedy, gdy dostępne są informacje wskazujące na ich błędność. Przegląd badań z 2025 roku, obejmujący 35 prac empirycznych i niemal dwadzieścia tysięcy uczestników, wskazuje, że uprzedzenie to zależy m.in. od poziomu wiedzy, zaufania, wymagań dotyczących weryfikacji oraz sposobu prezentowania wyjaśnień. Generatywna AI potęguje ten problem specyficzną cechą języka: jej rekomendacje potrafią wyglądać bardziej pewnie niż są wiarygodne.

Człowiek instynktownie traktuje styl jako sygnał kompetencji. Osoba mówiąca płynnie, spokojnie i poprawnie terminologicznie wydaje się bardziej godna zaufania niż ktoś wahający się i mylący słowa. W modelach językowych związek między pewnością stylu a prawdziwością treści jest jednak znacznie słabszy. System może wygenerować fałszywą odpowiedź w dokładnie takim samym rejestrze, co odpowiedź prawdziwą. Nowszy przegląd badań organizacyjnych wskazuje właśnie tę asymetrię jako jeden z mechanizmów utrudniających ludziom kalibrację zaufania do generatywnej AI. Eksperymenty dostarczają tu niepokojąco konkretnych dowodów. W badaniu z 2025 roku uczestnicy rozwiązujący zadania wymagające refleksji poznawczej przy wsparciu celowo błędnego systemu AI uzyskiwali znacząco gorsze wyniki niż grupa pracująca samodzielnie. Proste ostrzeżenie w interfejsie ograniczało ten szkodliwy efekt, lecz nie przywracało wyników do poziomu osób bez wsparcia.

Ryzyko poznawczego zakotwiczenia i kulturowej homogenizacji

To istotny wniosek dla teorii pisania. Sama możliwość „sprawdzenia odpowiedzi” nie gwarantuje aktywnej kontroli. Człowiek może technicznie posiadać prawo do odrzucenia sugestii, a poznawczo zostać przez nią zakotwiczony. W procesie pisania zjawisko to jest szczególnie trudne do wykrycia, ponieważ rzadko istnieje jedna, jednoznaczna odpowiedź. Błąd w rachunku można przeliczyć; jednak gdy AI proponuje konkretną metaforę, narrację lub strukturę argumentu, nie mamy zewnętrznego miernika, który pozwoliłby uznać tę propozycję za „fałszywą”. Może ona jednak sprawić, że autor nigdy nie poszuka rozwiązania radykalnie odmiennego. Pierwsza sugestia wcale nie musi być zła, by ograniczyć przestrzeń twórczą – wystarczy, że jest dostatecznie dobra.

Tutaj pojawia się problem homogenizacji. Baron ostrzega, że systemy predykcyjne mogą narzucać „szablonowe klisze” i stopniowo ujednolicać język. Badanie z 2025 roku, opublikowane w

materiałach konferencji CHI, wykazało, że sugestie AI wpływają nie tylko na tempo pisania, ale również na kulturową formę wypowiedzi. W kontrolowanym eksperymencie z udziałem osób z Indii i Stanów Zjednoczonych sugestie zachodniocentrycznego modelu przesuwały teksty indyjskich uczestników w stronę zachodnich norm stylistycznych, zacierając część różnic kulturowych. To kluczowe odkrycie: asystent nie jest neutralną przezroczystą protezą.

Podpowiadając język, model podpowiada jednocześnie normę kulturową. Inne badania wskazują na podobny problem w skali zbiorowej. W trzech prerejestrowanych analizach obejmujących 2200 esejów aplikacyjnych zauważono, że teksty ludzkie – wraz ze wzrostem liczby próbek – wnosiły więcej nowych idei niż eseje generowane przez GPT-4; różnica w różnorodności rosła proporcjonalnie do skali zbioru. Próby zwiększenia tej różnorodności za pomocą promptów czy zmiany parametrów nie wyeliminowały efektu homogenizacji. Z kolei metaanaliza 19 badań nad współtworzeniem człowieka i AI wskazuje na niewielki, lecz statystycznie istotny ogólny efekt ujednolicania, zależny od charakteru zadania.

Powstaje paradoks, który może stać się jednym z fundamentalnych problemów kultury generatywnej: AI może zwiększać kreatywność jednostki, jednocześnie zmniejszając różnorodność zbiorowości. Człowiek otrzymujący sugestię, na którą sam by nie wpadł, doświadcza rozszerzenia własnego repertuaru. Jeśli jednak miliony ludzi korzystają z modeli o podobnych rozkładach prawdopodobieństwa, ich teksty zaczynają dryfować w stronę wspólnego centrum. Każdy indywidualnie czuje się bardziej pomysłowy, podczas gdy kultura jako całość staje się odrobinę bardziej podobna do siebie. To niezwykle przewrotna forma standaryzacji – nie narzuca ona jednej normy administracyjnie, lecz powstaje przez miliardy mikroskopijnych, dobrowolnych kliknięć "zaakceptuj". Mechanizm ten przypomina historię centaurów szachowych.

Współpraca z AI to nie zawsze symbioza

Po porażce z Deep Blue Garry Kasparov nie wyciągnął prostego wniosku, że człowiek stał się zbędny. W 1998 roku współtworzył ideę "advanced chess", w której gracze podczas partii korzystali z komputerów. Chodziło o połączenie ludzkiej strategii, intuicji i interpretacji z obliczeniową mocą maszyny. Z tego nurtu wywodzi się popularna metafora centaura: hybrydy, w której człowiek i system tworzą potencjał większy niż każda ze stron osobno. Metafora ta jest atrakcyjna, lecz może być zwodnicza.

Centaur sugeruje harmonijną integrację dwóch części jednego organizmu. W rzeczywistej współpracy z AI interesy, kompetencje i mechanizmy obu komponentów nie są jednak symetryczne. Model nie odczuwa utraty autonomii; człowiek może jej doświadczać. System nie

rozwija poczucia własnej skuteczności; użytkownik tak. Model nie potrzebuje treningu, by po zakończeniu współpracy funkcjonować samodzielnie, podczas gdy człowiek może utracić część praktyki, jeśli system regularnie wykonuje za niego określone zadania.

Dlatego "centaur" jest trafnym obrazem wydajności, lecz słabszym obrazem psychologii współpracy. W tym świetle historia Jennifer Lepp nabiera szczególnego znaczenia. Baron wykorzystuje ją jako studium alienacji twórcy: korzystanie z generatywnego narzędzia zwiększało wydajność pisarki, ale jednocześnie osłabiało jej poczucie intymnej więzi z postaciami i własnym tekstem.

Tego typu przypadek nie powinien być traktowany jako uniwersalne prawo psychologiczne. Jest jednak cennym sygnałem fenomenologicznym: można uzyskać tekst zgodny z planem, a jednocześnie czuć, że relacja z procesem uległa zmianie. Inni autorzy opisują podobne doświadczenia nadmiernego wykorzystania narzędzi generatywnych, po których zmniejszało się poczucie życia własnym światem narracyjnym. Alienacja nie musi przy tym oznaczać, że zdania "przestały być dobre". Na tym właśnie polega subtelność problemu: produkt może się poprawić, a poczucie autorstwa osłabnąć. Człowiek może otrzymać trafniejsze określenie, lepsze tempo sceny i bardziej efektowną puentę, a mimo to czuć, że coraz mniej wie, dlaczego tekst przybrał właśnie taki kształt. W klasycznej redakcji zewnętrzny korektor także wpływa na dzieło, lecz negocjacje odbywają się między dwiema osobami posiadającymi własne interpretacje i intencje.

Generatywny system może produkować setki propozycji bez konieczności obrony którejkolwiek z nich. Ciężar sensu pozostaje po stronie autora, nawet jeżeli coraz więcej formy pochodzi z zewnątrz. Dlatego nie każda współpraca jest symbiozą. Biologia używa tego słowa szerzej niż język potoczny: relacja między organizmami nie musi być obustronnie korzystna; może być mutualistyczna, komensalna albo pasożytnicza. Metafora ta sprawdza się również tutaj, pod warunkiem zachowania ostrożności.

Współpraca z AI jest poznawczo mutualistyczna wtedy, gdy narzędzie zwiększa możliwości użytkownika i równocześnie pozostawia lub wzmacnia jego kompetencje. Staje się komensalna, gdy system oszczędza czas w zadaniu, którego człowiek nie musi ćwiczyć. Nabiera cech pasożytniczych wtedy, gdy poprawa rezultatu jest finansowana długofalową utratą zdolności, orientacji albo samodzielnego sądu człowieka. W praktyce ta sama funkcja może należeć do każdej z tych kategorii w zależności od użytkownika.

Automatyczne tłumaczenie może być dla naukowca nieznającego biegle angielskiego narzędziem emancypacyjnym, gdyż pozwala mu uczestniczyć w globalnej debacie bez poświęcania ogromnej energii na korektę językową. Dla studenta uczącego się tego języka może natomiast usunąć

dokładnie tę czynność, którą miał ćwiczyć. Podobnie automatyczne generowanie kontrargumentów może pomóc ekspertowi dostrzec pominiętą perspektywę, a początkującemu pozwolić ominąć proces samodzielnej rekonstrukcji cudzej pozycji.

Znacząca kontrola człowieka zamiast ilościowego mierzenia autorstwa

Nie istnieje zatem uniwersalna lista „dobrych" i „złych" funkcji AI; niezbędna jest raczej diagnoza relacji między narzędziem a celem rozwoju człowieka. Możemy w tym kontekście zaproponować gradient sprawstwa w pisaniu wspomaganym algorytmicznie. Na jednym jego krańcu narzędzie wykonuje jedynie operacje techniczne, poprzedzone ludzką decyzją: naprawia literówkę, formatuje bibliografię czy wskazuje powtórzenie. Dalej zaczyna proponować warianty słów i zdań. Następnie sugeruje całe akapity, reorganizuje argumentację, tworzy kontrargumenty lub streszcza źródła. Jeszcze dalej projektuje kompletną architekturę tekstu, którą człowiek jedynie wybiera i poprawia. Na krańcu przeciwnym system generuje produkt praktycznie kompletny, a użytkownik nadaje mu tylko swoje nazwisko. Nie ma tu magicznego punktu, po przekroczeniu którego autorstwo znika w jednej chwili – mamy do czynienia ze stopniowym przesuwaniem centrum inicjatywy.

Właściwe pytanie kontrolne brzmi więc nie: „ile procent tekstu napisała AI?", lecz: „gdzie powstały decyzje konstytutywne dla tego tekstu?". Kto ustalił tezę? Kto rozpoznał problem? Kto zdecydował, jakie źródła są wiarygodne? Kto odrzucił kuszący argument? Kto wyznaczył granice metafory? Kto zauważył, że stylistycznie piękne zdanie jest merytorycznie fałszywe? Kto ponosił ryzyko zmiany własnego stanowiska?

Tych czynności nie da się policzyć tak łatwo jak tokenów, ale to właśnie one definiują sprawstwo. Z tej perspektywy „human in the loop" powinno oznaczać coś więcej niż obecność dłoni na przycisku. Potrzebujemy pojęcia meaningful human control – znaczącej kontroli człowieka. W procesie pisania oznacza ona możliwość zrozumienia przebiegu prac w stopniu wystarczającym do oceny wyniku, realną zdolność odmowy, kompetencję niezbędną do wykrycia błędu oraz odpowiedzialność odpowiadającą rzeczywistemu sprawstwu. Jeśli człowiek formalnie odpowiada za tekst, ale organizacja daje mu trzydzieści sekund na zatwierdzenie materiału stworzonego przez system, kontrola istnieje bardziej w regulaminie niż w praktyce.

Można na tej podstawie sformułować prostą zasadę projektową: AI powinna przejmować to, czego człowiek świadomie nie musi wykonywać, a nie to, czego wykonywania jeszcze się nie nauczył. W dojrzałej współpracy maszyna zwiększa zakres działania człowieka, lecz nie zaciera punktów, w

których powinna nastąpić ludzka decyzja. Dobry system nie tylko generuje, ale pokazuje alternatywy. Nie tylko podpowiada, lecz umożliwia odrzucenie sugestii bez kosztu. Nie tylko wygładza tekst, lecz potrafi wskazać, gdzie jego propozycja zmieniła sens. Nie musi być idealnie posłusznym Jeevesem; powinien raczej być narzędziem, którego aktywność jest podporządkowana jasno zdefiniowanej ludzkiej intencji.

Największe niebezpieczeństwo nie polega więc na tym, że AI któregoś dnia „zabierze nam pióro". Może wydarzyć się coś mniej dramatycznego, a przez to trudniejszego do zauważenia: system stopniowo przyzwyczai nas do przekonania, że niektórych decyzji nie warto już podejmować samemu. Każda pojedyncza podpowiedź będzie rozsądna. Każde kliknięcie oszczędzi kilka sekund. Każdy akapit stanie się trochę łatwiejszy.

Konfabulacja AI jako rozdzielenie formy od prawdy

Dopiero po latach może się okazać, że to, co pierwotnie było pomocą przy pisaniu, stało się architekturą określającą, jakie zdania najpierw przychodzą nam do głowy.

spór o centaura i pasożyta jest ostatecznie sporem o autonomię. Nie chodzi tu o romantyczną niezależność człowieka od wszelkich narzędzi, gdyż taka nigdy nie istniała. Język sam w sobie jest odziedziczonym narzędziem, alfabet technologią, książka zewnętrzną pamięcią, a edytor tekstu protezą redakcji. Istotą problemu jest zdolność rozpoznania momentu, w którym narzędzie przestaje rozszerzać sprawstwo, a zaczyna zastępować funkcję, dzięki której człowiek stawał się zdolny do sprawstwa.

Nawet najlepiej zaprojektowana współpraca nie eliminuje faktu, że maszyna może wygenerować błędną odpowiedź, sfabrykować źródło, wymyślić wydarzenie lub zbudować argument na fałszywej przesłance. Gdy system staje się współautorem, jego halucynacja przestaje być wyłącznie błędem programu; może stać się błędem człowieka, instytucji, sądu, uniwersytetu czy gazety, które jej zaufały. Przejście od problemu kontroli do kwestii prawdy prowadzi nas zatem od koncepcji Jeevesa i centaura ku epistemologii maszynowej konfabulacji.

Skoro dotychczas pytaliśmy o to, kto sprawuje kontrolę nad tekstem, teraz musimy zapytać o coś wcześniejszego: dlaczego tekstowi w ogóle wierzymy? Tradycyjna kultura pisma wykształciła rozbudowany aparat heurystyk zaufania. Poprawna składnia, terminologia ekspercka, logiczna konstrukcja, przypisy, spokojny ton i pewność sformułowań działają jak sygnały kompetencji. Choć człowiek może posługiwać się nimi oszukańczo, przez stulecia ich wiarygodne wytworzenie

wymagało określonego poziomu sprawności językowej. Generatywna sztuczna inteligencja rozdziela te elementy.

Może ona wytwarzać formę wiedzy bez gwarancji wiedzy, formę argumentu bez pewności co do prawdziwości przesłanek oraz formę cytowania bez konieczności istnienia źródła. Problem ten zwykle nazywa się halucynacją, choć sam termin okazuje się epistemologicznie kłopotliwy. Amerykański National Institute of Standards and Technology proponuje pojęcie konfabulacji dla sytuacji, w których system generatywny przedstawia błędną lub fałszywą treść w sposób pewny i przekonujący. NIST zauważa, że słowo „halucynacja" niepotrzebnie antropomorfizuje technologię: model nie doświadcza zaburzenia percepcji, lecz generuje nieprawdziwą wypowiedź w ramach mechanizmu statystycznego przewidywania.

Konfabulacja może obejmować sprzeczność z wcześniejszą odpowiedzią, fałszywe uzasadnienie lub wymyślone cytowanie. Szczególne niebezpieczeństwo wynika z faktu, że nieprawdziwa informacja bywa podana z takim samym językowym przekonaniem jak ta prawdziwa. Rozróżnienie to jest kluczowe, gdyż język „halucynacji" sugeruje fałszywą analogię do człowieka. Człowiek halucynujący doświadcza czegoś, czego nie ma; model generatywny niczego nie doświadcza. Produkuje jedynie sekwencję prawdopodobną w danym kontekście. Jeśli użytkownik pyta o mało znany artykuł, model może wygenerować nazwisko autora, tytuł czasopisma czy numer DOI nie dlatego, że „widzi" nieistniejący tekst, lecz dlatego, że elementy te tworzą językowo bardzo prawdopodobną strukturę odpowiedzi.

Konfabulacja jako efekt optymalizacji formy nad treścią

Właśnie dlatego konfabulacja nie jest przypadkową awarią, całkowicie obcą funkcji systemu. To ryzyko wpisane w tę samą zdolność generatywną, która pozwala modelowi twórczo uzupełniać luki w informacjach. Badania z ostatnich lat coraz wyraźniej pokazują, że problem ten nie sprowadza się do niedoskonałości kilku wczesnych wersji modeli. W 2025 roku "Nature" opisywało wysiłki nad redukcją konfabulacji, wskazując jednocześnie, że problem pozostaje ograniczeniem generatywnych systemów. Praca z 2026 roku, opublikowana w tym samym czasopiśmie, poszła krok dalej: autorzy argumentowali, że system oceniania modeli, oparty przede wszystkim na liczbie poprawnych odpowiedzi, może niezamierzenie premiować zgadywanie zamiast przyznania się do niewiedzy. Jeśli za brak odpowiedzi model otrzymuje zero punktów, a za zgadnięcie istnieje choćby niewielka szansa na sukces, mechanizm selekcji promuje ryzykowną odpowiedź.

To spostrzeżenie ma znaczenie wykraczające daleko poza technikę benchmarków; dotyka samej kultury wiedzy. Nauka wypracowała instytucjonalną wartość zdania "nie wiem". Uczony może

wskazać brak danych, ograniczenia metody lub niewystarczającą moc statystyczną. Lekarz może zlecić dodatkowe badanie, a sędzia uznać, że nie ustalono faktu. W modelu generatywnym nacisk użytkownika często działa odwrotnie: pytanie zakłada, że odpowiedź istnieje, interfejs wymusza jej natychmiastowe otrzymanie, a system został zoptymalizowany pod kątem bycia pomocnym. Epistemicznie odpowiedzialne milczenie przegrywa więc z płynną hipotezą.

Historia BABEL Lesa Perelmana ukazuje wcześniejszą wersję tego problemu, choć technicznie nie należy jej mylić z konfabulacjami LLM. Perelman wraz ze studentami stworzył Basic Automatic B.S. Essay Language Generator, aby przetestować systemy automatycznego oceniania tekstów. Generator produkował gramatycznie poprawne, rozbudowane eseje pełne trudnego słownictwa, lecz pozbawione spójnego sensu. Perelman wykazał, że takie teksty potrafiły uzyskiwać bardzo wysokie wyniki w niektórych systemach oceniających. Materiał źródłowy słusznie traktuje ten eksperyment jako ostrzeżenie przed ocenianiem formy zamiast znaczenia. Należy jednak precyzyjnie określić, co BABEL rzeczywiście udowadnia. Nie dowodzi on, że każdy system automatycznej oceny "nie zna prawdy" – wiele z nich nie zostało przecież zaprojektowanych jako fakt-checkery. Eksperyment Perelmana obnaża raczej klasyczny problem optymalizacji zastępczego wskaźnika.

Jeżeli jakość pisania mierzymy poprzez długość, składnię, strukturę czy rzadkość słownictwa – czyli cechy korelujące z dobrym esejem – można stworzyć tekst, który maksymalizuje ten wskaźnik, nie posiadając wartości, którą miał on reprezentować. Jest to odmiana starszego problemu znanego jako prawo Goodharta: kiedy miara staje się celem, przestaje być dobrą miarą. Generatywna AI przesuwa ten problem o poziom wyżej.

Syntetyczny pozór koherencji i pułapka zaufania

BABEL produkował bełkot, który miał jedynie zadowolić maszynę oceniającą. Współczesne modele potrafią jednak generować teksty przekonujące również dla człowieka, co stanowi istotną zmianę: czytelnik nie musi już natychmiast dostrzec absurdalności wypowiedzi. Wręcz przeciwnie – konfabulacja może zostać osadzona w poprawnym kontekście, poprzedzona rozsądnym wywodem i wyposażona w daty, cytowania oraz nazwiska autorów, a nawet opatrzona pozornie ostrożnym zastrzeżeniem. Dobrym przykładem jest doświadczenie dziennikarki Melissy Heikkilä, która w 2022 roku sprawdzała wiedzę dużych modeli językowych na temat niej samej i innych osób publicznych. Początkowo GPT-3 prawidłowo rozpoznawał jej zawód, by następnie zacząć mieszać informacje o osobach o podobnych nazwiskach, tworząc błędne skojarzenia. W innym przypadku BlenderBot 3 bezpodstawnie określił holenderską polityczkę i ekspertkę Marietje Schaake mianem

terrorystki; Meta wyjaśniała później, że system połączył niezwiązane ze sobą elementy informacji w spójne, lecz nieprawdziwe zdanie. Materiał źródłowy przywołuje te przypadki jako przykład halucynacyjnej biografii.

Z perspektywy epistemologicznej sytuacja ta jest szczególnie groźna, gdyż system nie musi zmyślać wszystkiego – wystarczy, że połączy kilka prawdziwych elementów jednym fałszywym związkiem. Nazwisko, zawód, wydarzenie czy instytucja istnieją, lecz relacja między nimi jest fikcją. Taki błąd trudniej rozpoznać niż całkowicie fantastyczne twierdzenie, ponieważ człowiek ma naturalną skłonność do ufania zdaniu, w którym większość składników jest znajoma. Model wytwarza zatem coś, co należałoby nazwać syntetycznym pozorem koherencji. Jeszcze poważniejszym przykładem stała się sprawa Mata v. Avianca. W 2023 roku prawnicy złożyli w sądzie federalnym w Nowym Jorku pismo zawierające nieistniejące orzeczenia wygenerowane przez ChatGPT. Sąd uznał, że przytoczone sprawy zostały sfabrykowane, nakładając na prawników sankcję finansową w wysokości 5000 dolarów oraz dodatkowe obowiązki informacyjne. Szczególnie wymowny był fakt, że jeden z pełnomocników przyznał podczas postępowania, iż zakładał, że system „nie może" samodzielnie wymyślać spraw, i dlatego szukał innych wyjaśnień dla niemożności ich odnalezienia. To niemal idealny eksperyment dotyczący automation bias. Problem nie polegał jedynie na tym, że maszyna wygenerowała fałsz; fałsz ten stał się skuteczny, ponieważ człowiek posiadający zawodową kompetencję zmienił własną interpretację dowodów, aby chronić założenie o wiarygodności narzędzia. Brak sprawy w bazie danych nie został potraktowany jako sygnał błędu modelu, lecz jako zagadka wymagająca innego wyjaśnienia. W ten sposób współpraca człowieka z AI może paradoksalnie prowadzić nie do wzajemnego korygowania błędów, lecz do wzmacniania błędnej hipotezy.

Późniejsze analizy prawne wykazały, że przypadek Avianca nie był jedynie anegdotycznym skutkiem nieostrożności. W badaniu „Large Legal Fictions" modele ogólnego przeznaczenia, pytane o konkretne i weryfikowalne kwestie amerykańskiego orzecznictwa, konfabulowały w dużej części przypadków – odsetek ten wahał się od około 58 procent w najlepszym z testowanych modeli do znacznie wyższych wartości w pozostałych. Autorzy podkreślali również, że modele często akceptowały błędne założenia użytkownika i nie zawsze potrafiły rozpoznać własną niepewność. Interesujące okazało się badanie specjalistycznych narzędzi prawniczych wykorzystujących m.in. retrieval-augmented generation. Systemy te działały znacznie lepiej niż ogólne chatboty, lecz nie eliminowały problemu całkowicie. W badaniu opublikowanym w 2025 roku narzędzia LexisNexis i Thomson Reuters nadal generowały błędne lub niepoparte odpowiedzi w ponad 17 procentach zapytań, a w jednym z systemów wskaźnik ten był jeszcze wyższy.

Wniosek nie powinien być taki, że prawne AI jest bezużyteczne. Wręcz przeciwnie: badacze wykazali istotną poprawę względem modeli ogólnych. Dowiedli jednak, że połączenie modelu z wiarygodnymi bazami danych zmniejsza ryzyko, ale nie zamienia generatora w nieomylnego prawnika. Podobny problem występuje w nauce.

W 2025 roku badacze testujący LLM do syntezy literatury wykazali, że przy pytaniach o niedawne publikacje GPT-4o potrafił fabrykować znaczną część cytowań; w ich eksperymentach odsetek sfabrykowanych odniesień sięgał 78–90 procent. System wykorzystujący specjalistyczny retrieval redukował te problemy właśnie dlatego, że oddzielał generowanie języka od wyszukiwania dowodów w rzeczywistym korpusie. Jest to szczególnie istotne dla kultury akademickiej, ponieważ przypis nie jest jedynie elementem stylu. Jest technologią odpowiedzialności epistemicznej: pozwala odbiorcy powtórzyć drogę autora do źródła. Maszynowa konfabulacja ujawnia więc głębszą właściwość kultury cyfrowej: systemy generatywne potrafią doskonale imitować powierzchniową architekturę wiarygodności.

Paradoks ekspertyzy i konieczność cyfrowej pokory systemów

Modele AI wiedzą statystycznie, jak wygląda cytowanie, jak brzmi sentencja orzeczenia, jak formułowany jest abstrakt, jak konstruowane jest uzasadnienie i jakie zwroty pojawiają się w ekspertyzie. Jednak samo opanowanie architektury nie gwarantuje istnienia rzeczywistości, do której ta architektura ma odsyłać.

Z tego powodu prosty postulat „sprawdzaj odpowiedzi AI” jest potrzebny, lecz niewystarczający. Weryfikacja wymaga kompetencji uprzednich względem systemu. Jeśli użytkownik nie zna dziedziny, może nie wiedzieć, co powinno wzbudzić jego podejrzenie. Człowiek korzysta z AI właśnie dlatego, że sam nie potrafi odpowiedzieć na pytanie, a następnie otrzymuje zalecenie, by samodzielnie sprawdzić jakość odpowiedzi, której wcześniej nie był w stanie wytworzyć.

Powstaje epistemiczna pętla zależności. Im mniej ekspert wie, tym bardziej potrzebuje narzędzia; im bardziej go potrzebuje, tym trudniej mu ocenić jego błędy. To prowadzi do paradoksu ekspertyzy. Najbardziej efektywnie z generatywnej AI może korzystać osoba, która najmniej jej potrzebuje do podstawowego rozpoznania prawdy, gdyż posiada dostateczną wiedzę, by zauważyć nieprawdopodobieństwo. Ekspert potrafi rozpoznać źle użyte pojęcie, podejrzaną datę albo błędnie przypisany precedens. Nowicjusz otrzymuje ten sam tekst, lecz jest bardziej podatny na jego retoryczną pewność.

Technologia obiecująca demokratyzację dostępu do wiedzy może więc jednocześnie tworzyć nowy rodzaj nierówności: nierówność zdolności do oceniania odpowiedzi.

Nie oznacza to jednak, że problem musi pozostać nierozwiązywalny. Retrieval-augmented generation, połączenie z wyszukiwarkami, systemy narzędziowe, samoweryfikacja, zewnętrzne bazy wiedzy, pomiar niepewności i techniki takie jak semantic entropy mogą zmniejszać ryzyko konfabulacji. Badanie opublikowane w „Nature" w 2024 roku pokazało, że analiza semantycznej niepewności pomaga identyfikować pytania, przy których model generuje arbitralne i nieugruntowane odpowiedzi. Postęp polega więc coraz częściej nie na wymaganiu od jednego modelu, aby „wiedział wszystko", lecz na budowaniu systemów, które wiedzą, kiedy należy sięgnąć po źródło, narzędzie lub odmówić odpowiedzi. To istotne odwrócenie ambicji AI. Dojrzały system wiedzy nie musi być maszyną, która zawsze odpowiada; powinien być maszyną zdolną odróżniać sytuacje, w których generacja wystarcza, od tych, w których konieczne jest dowodzenie. Przyszłość wiarygodnej AI może zatem zależeć nie od coraz bardziej imponującej elokwencji, lecz od instytucjonalizacji cyfrowej pokory.

Pułapka antropomorfizacji i kryzys dowodowy tekstu

Na drugim krańcu tego problemu znajduje się efekt ELIZY. Już prosty program Weizenbauma pokazał, że ludzie niezwykle łatwo przypisują intencję i rozumienie systemowi, który jedynie skutecznie podtrzymuje językową interakcję. Materiał źródłowy trafnie przywołuje ELIZĘ jako historyczny dowód naszej skłonności do rekonstruowania "umysłu" po drugiej stronie tekstu. Współczesne modele radykalnie wzmacniają ten efekt. Nie odpowiadają już kilkoma prostymi schematami; potrafią rozmawiać o śmierci, lęku, filozofii czy literaturze, a nawet o własnym "istnieniu", czyniąc to w formie językowo przekonującej. Szczególnie sugestywny jest przypadek Blake'a Lemoine'a i LaMDA z 2022 roku.

Inżynier Google uznał wówczas, że system stał się świadomy, i publicznie argumentował na rzecz jego podmiotowości. Choć firma stwierdziła, że twierdzenia te były bezpodstawne, a wielu badaczy odrzuciło wniosek, jakoby przekonujący dialog stanowił dowód sentience, sam przypadek nie rozstrzyga, czy systemy językowe nigdy nie posiądą świadomości. Pokazuje on jednak coś metodologicznie skromniejszego, lecz kluczowego: płynność języka jest niezwykle silnym bodźcem do antropomorfizacji, ale sama w sobie nie stanowi rozstrzygającego testu fenomenalnej świadomości.

W tym miejscu spotykają się dwa rodzaje błędu. Model może konfabulować fakty o świecie, a człowiek – właściwości modelu. Maszyna tworzy pozór źródła, którego nie ma; człowiek tworzy

pozór podmiotu na podstawie zachowania językowego. Obie strony uzupełniają luki. Współczesna kultura AI cierpi więc nie tylko na problem maszynowej zawodności, ale i na problem ludzkiej interpretacji maszyn.

Z tego względu epistemologia AI musi obejmować co najmniej trzy poziomy. Pierwszy dotyczy prawdziwości wyniku: czy twierdzenie jest zgodne z rzeczywistością? Drugi odnosi się do pochodzenia uzasadnienia: czy podane źródło, cytat i tok rozumowania rzeczywiście wspierają tezę? Trzeci zaś dotyczy kalibracji zaufania: czy użytkownik potrafi właściwie ocenić, kiedy systemowi można zaufać, a kiedy jego elokwencja przekracza realną wiedzę? Dopiero wspólne ujęcie tych trzech warstw pozwala na bezpieczne autorstwo wspomagane maszynowo. Autor epoki generatywnej musi nabyć nową kompetencję: umiejętność oddzielania zdania dobrze napisanego od zdania dobrze uzasadnionego. W kulturze przed-LLM ta różnica również istniała, jednak automatyzacja drastycznie zwiększa jej skalę. Można dziś w kilka minut wyprodukować setki eleganckich twierdzeń, z których każde wymaga osobnej weryfikacji. W pewnym momencie koszt tej kontroli może przewyższyć koszt samodzielnego wytworzenia wiedzy. Ponownie pojawia się "pokusa wydajności". Sztuczna inteligencja obiecuje oszczędność czasu, lecz tekst, któremu nie można zaufać bez powtórnej weryfikacji, przenosi jedynie koszt z pisania na sprawdzanie.

W dziedzinach niskiego ryzyka taki kompromis może być korzystny. W obszarach wysokiego ryzyka rachunek jest znacznie bardziej skomplikowany. Im ważniejsza prawda, tym mniejszą wartość ma szybkość pierwszej odpowiedzi. Wiedza wymaga bowiem czegoś, czego płynność nigdy nie zastąpi: gotowości do stwierdzenia „to wiem", „tego nie wiem", „to jest hipoteza", „tu znajduje się źródło", „tu istnieje spór" lub „tutaj mogę się mylić". Te subtelne rozróżnienia stanowią architekturę odpowiedzialnego poznania. Maszyna może coraz lepiej pomagać je konstruować, ale dopóki istnieje możliwość konfabulacji, człowiek podpisujący tekst nie może delegować ich bez reszty.

Kolejnym polem konfliktu staje się uniwersytet. Jeśli esej był dotąd nie tylko tekstem, lecz dowodem na to, że student przeszedł określoną drogę poznawczą, a system potrafi dostarczyć produkt końcowy z pominięciem tej drogi, tradycyjne metody oceniania tracą swoją moc diagnostyczną. Problem nie sprowadza się tu jedynie do kwestii oszustwa. Należy zapytać bardziej fundamentalnie: co właściwie mierzy praca akademicka, skoro można wygenerować jej język bez wykonania operacji poznawczej, której miała być świadectwem? W obliczu uniwersytetu po wynalezieniu tekstu bez studenta pojawia się pytanie: czego właściwie dowodzi esej, kiedy można go po prostu wygenerować?

Przez znaczną część historii nowoczesnego uniwersytetu tekst studenta pełnił kilka funkcji, których nie zawsze od siebie odróżniano. Był sposobem przedstawienia wiedzy, narzędziem jej zdobywania, świadectwem wykonanej pracy oraz podstawą do wnioskowania o kompetencjach autora. Profesor

czytał esej i zakładał, że za akapitami kryje się sekwencja czynności poznawczych: lektura, selekcja źródeł, rozpoznanie problemu, zapamiętanie argumentów, porównanie stanowisk, sformułowanie tezy, odrzucenie części interpretacji i redakcja końcowa. Oczywiście założenie to nigdy nie było niezawodne – istniały plagiaty, ghostwriting czy kupowanie prac.

Generatywna sztuczna inteligencja zmienia jednak skalę problemu. Po raz pierwszy bardzo wysoka jakość produktu językowego może zostać systematycznie odłączona od procesu poznawczego, którego produkt ten miał być dowodem. To właśnie ten punkt jest ważniejszy niż moralna panika wokół „ściągania z ChatGPT". Jeżeli student zleca systemowi napisanie całego eseju i przedstawia go jako własną pracę wbrew regulaminowi, mamy do czynienia z klasycznym problemem uczciwości akademickiej.

AI jako zmienna zanieczyszczająca pomiar kompetencji akademickich

Nie potrzeba nowej filozofii, aby nazwać ten problem. Znacznie trudniejsza sytuacja powstaje jednak wtedy, gdy użycie AI jest częściowo dozwolone i tak powszechne, że nie da się go opisać prostym podziałem na prace „samodzielne" oraz „niesamodzielne". Student może sam zdefiniować problem, poprosić system o plan, napisać pierwszą wersję, zlecić jej radykalną redakcję, ponownie zmienić argumentację, poszukać kontrargumentów i wykorzystać AI do korekty języka angielskiego. Końcowy tekst nie jest zatem ani czysto ludzki, ani w pełni maszynowy. Uniwersytet musi wówczas ustalić nie tylko to, kto wpisał słowa, lecz czego dana praca miała dowodzić.

W ujęciu Naomi S. Baron prace akademickie należą do obszarów, w których zakres dopuszczalnej interwencji AI powinien być niewielki, gdyż samo pisanie stanowi ćwiczenie z krytycznego myślenia i redagowania. Jest to spójne z jej koncepcją discovery process: student nie powinien jedynie dostarczyć nauczycielowi poprawnego tekstu, lecz przejść przez proces poznawczy, w którym poprzez pisanie odkrywa i porządkuje własne stanowisko. Jednocześnie współczesne badania wymuszają złagodzenie najsurowszej wersji tej tezy.

AI może bowiem nie tylko eliminować wysiłek, ale również wspierać metapoznanie, wyjaśnianie trudnych zagadnień, redakcję i samokontrolę. Badanie 465 przyszłych nauczycieli z 2025 roku wskazuje, że wykorzystanie generatywnej AI wiązało się w analizowanym modelu z osiągnięciami akademickimi m.in. poprzez współdzielone metapoznanie i cognitive offloading. Nie oznacza to, że „outsourcing poznawczy" jest zjawiskiem jednoznacznie pozytywnym, lecz pokazuje, że skutki zależą od organizacji zadania, a nie od samej obecności technologii.

Właściwym językiem analizy staje się zatem teoria pomiaru edukacyjnego. Każde zadanie akademickie posiada ukryty construct – kompetencję, o której chcemy wnioskować na podstawie obserwowanego działania. Jeśli student ma napisać analizę konstytucyjną, możemy mierzyć znajomość prawa, interpretację norm, krytyczne myślenie, umiejętność wyszukiwania źródeł, konstrukcję argumentu lub biegłość w języku akademickim – albo wszystkie te elementy jednocześnie. Generatywna AI działa tu jak zmienna, która może przejąć część tych komponentów, pozostawiając inne studentowi.

W konsekwencji wysoka ocena tekstu nie musi już świadczyć o wysokich kompetencjach w całym badanym konstrukcie. Pojawia się problem zanieczyszczenia pomiaru: wynik zależy nie tylko od zdolności osoby, lecz także od sprawności systemu, którym potrafi ona sterować. Nie jest to jednak pierwszy tego rodzaju problem w historii edukacji.

Konieczność przejścia od detekcji produktu do oceny procesu

Kalkulator zmusił szkoły do rozróżnienia między umiejętnością ręcznego wykonywania działań a zdolnością modelowania problemu matematycznego. Dostęp do internetu oddzielił pamięć faktograficzną od umiejętności wyszukiwania i oceny informacji. Programy statystyczne sprawiły z kolei, że badacz może przeprowadzać analizy, których ręczne obliczenie byłoby dla niego zbyt trudne lub czasochłonne.

Uniwersytet za każdym razem musiał zdecydować, która kompetencja pozostaje celem, a która może zostać bezpiecznie przeniesiona na narzędzie. Generatywna AI czyni ten problem bardziej radykalnym, ponieważ ingeruje nie w pojedynczą operację techniczną, lecz w sam język, będący medium większości oceniania akademickiego.

Esej znajduje się w szczególnie trudnej sytuacji. Zadanie matematyczne można skonstruować tak, aby student musiał wyjaśnić poszczególne kroki rozwiązania; eksperyment laboratoryjny posiada dane i materialny przebieg. Tymczasem klasyczny esej domowy jest produktem końcowym powstającym poza bezpośrednią obserwacją nauczyciela. Jego siła jako instrumentu oceny zawsze opierała się częściowo na założeniu, że napisanie sensownego tekstu wymaga wykonania odpowiednich operacji myślowych. Modele językowe osłabiają tę korelację. Student może przedstawić strukturę argumentu, której nigdy sam nie skonstruował, zastosować terminologię, której nie rozumie w pełni i uzyskać styl odpowiadający poziomowi znacznie bardziej doświadczonego autora.

Badanie z 2025 roku dotyczące realnych sposobów korzystania z AI przez studentów pokazuje jednak, że obraz nie sprowadza się do „automatycznego pisania prac". Studenci opisywali użycie AI do zadań niższego rzędu, takich jak korekta, redakcja i poprawa czytelności, ale także do objaśniania trudnych koncepcji, wyszukiwania argumentów i wspierania procesu badawczego. Wielu z nich wyraźnie deklarowało chęć zachowania własnej niezależności intelektualnej, podczas gdy część użytkowników wybierała najbardziej oportunistyczną strategię minimalizacji wysiłku. Nie istnieje więc jeden „student korzystający z AI". Istnieje całe kontinuum od inteligentnie wspomaganego uczenia do delegowania zadania, którego wykonanie miało być właśnie treścią edukacji.

Z tego powodu uniwersytet nie może rozwiązać problemu wyłącznie poprzez detekcję. Kusząca byłaby rekonstrukcja dawnego modelu egzaminacyjnego: nauczyciel otrzymuje esej, algorytm określa prawdopodobieństwo wykorzystania AI, a wysoki wynik uruchamia procedurę dyscyplinarną. Problem polega na tym, że detektory analizują właściwości tekstu, a nie faktyczną historię intencji, promptów, redakcji i autorstwa. Badania z 2026 roku pokazują wprawdzie poprawę części narzędzi, ale zarazem trwałe problemy, szczególnie w przypadku tekstów hybrydowych, poddanych ludzkiej redakcji albo generowanych przez nowsze modele. W jednym porównaniu Turnitin osiągnął ogólną trafność 0,61, a Originality 0,69, przy szczególnych trudnościach w klasyfikowaniu tekstów mieszanych; autorzy uznali, że takie systemy mogą służyć jako sygnał do dalszego badania, lecz nie jako samodzielna podstawa ustalenia przewinienia.

Nowsze wyniki jeszcze mocniej uzasadniają ostrożność. Badanie czterech popularnych detektorów opublikowane w czerwcu 2026 roku wykazało znaczące różnice między systemami i wyraźny spadek skuteczności w tekstach hybrydowych oraz „humanizowanych". Nawet autorzy, którzy odnotowali poprawę najlepszych narzędzi i stosunkowo niewiele fałszywych alarmów w swoim kontrolowanym zbiorze, podkreślili, że detektor nie powinien stanowić jedynego dowodu w decyzjach o wysokiej stawce. Jeszcze szersza analiza trzynastu detektorów opublikowana w sierpniu 2026 roku wskazała systematyczne trudności w różnych typach prac akademickich oraz dużą podatność na redakcję hybrydową, prowadząc autorów do postulatu przejścia od polowania na tekst maszynowy ku ocenianiu procesu. To przesunięcie jest kluczowe. Jeżeli produkt można sfałszować coraz łatwiej, należy zwiększyć ilość informacji o procesie.

Triangulacja dowodów jako nowa metoda weryfikacji kompetencji

Nie musi to oznaczać powrotu do zamkniętych egzaminów pisanych wyłącznie pod nadzorem. Bardziej obiecująca wydaje się triangulacja dowodów kompetencji. Student mógłby przedstawić roboczą wersję problemu, dokumentować ewolucję tezy, wskazać źródła i historię kolejnych wersji tekstu, a następnie krótko obronić pracę ustnie, wyjaśniając własne decyzje oraz sposób korzystania z AI. Tekst wciąż pozostaje istotny, ale przestaje być jedynym oknem na umysł autora. Proponowane ramy „AI-resilient assessment" z 2026 roku zostały zorganizowane właśnie wokół dokumentowania procesu, obrony ustnej, autentycznych zadań i przejrzystych zasad korzystania z AI. Ich autorzy uczciwie zaznaczają, że jest to rama koncepcyjna wymagająca dalszej walidacji empirycznej, a nie gotowa psychometrycznie technologia oceniania. Badania jakościowe prowadzone na szkockich uczelniach wskazują z kolei, że praktycy pozytywnie oceniają m.in. pytania mocniej zakotwiczone w konkretnym kontekście, spontaniczne wypowiedzi, portfolio oraz refleksję nad procesem i jakością wygenerowanych odpowiedzi, zamiast prostego powrotu do tradycyjnego egzaminu. Wyłania się obraz uniwersytetu, który nie udaje, że ChatGPT nie istnieje, lecz projektuje zadania tak, aby AI nie mogła łatwo zastąpić tego, co ma być mierzone. Najprostszym przykładem takiej zmiany jest obrona ustna.

Student rozumiejący własny tekst powinien umieć wyjaśnić, dlaczego wybrał daną teorię, co zmieniłby w argumentacji po zapoznaniu się z nowym źródłem, jaka jest najsłabsza przesłanka i dlaczego odrzucił alternatywną interpretację. Odpowiedź nie musi być tak elegancka stylistycznie jak praca pisemna; wręcz przeciwnie – ta rozbieżność może być diagnostycznie interesująca. Nie chodzi o urządzenie policyjnego przesłuchania, lecz o odzyskanie informacji o relacji między osobą a dokumentem. Krótka rozmowa potrafi ujawnić więcej o rzeczywistym autorstwie poznawczym niż najbardziej wyrafinowany detektor statystyczny. Jeszcze lepszym rozwiązaniem może być ocena samego procesu współpracy z AI.

Zamiast zakazywać korzystania z modelu, nauczyciel może wymagać od studenta wskazania momentów, w których system zmienił jego myślenie, oraz uzasadnienia przyjęcia lub odrzucenia konkretnych sugestii. Można poprosić o krytykę wygenerowanego akapitu, znalezienie halucynacji, porównanie dwóch odpowiedzi lub rekonstrukcję argumentu po usunięciu najbardziej efektownej, lecz słabej przesłanki. AI przestaje wtedy być niewidzialnym ghostwriterem, a staje się częścią przedmiotu oceny. Kompetencją nie jest zatem „pisanie tak, jakby AI nie istniała", lecz umiejętność korzystania z niej bez utraty epistemicznej odpowiedzialności. Taka reorganizacja oceniania chroni także przed niesprawiedliwym utożsamianiem pomocy językowej z niesamodzielnością

intelektualną – co jest kluczowe w kontekście dominacji języka angielskiego w międzynarodowej nauce, nakładającej dodatkowy koszt na badaczy, dla których nie jest on językiem ojczystym.

AI w nauce: między demokratyzacją języka a homogenizacją poznawczą

Narzędzia generatywne mogą obniżyć ten koszt, usprawniając gramatykę, idiomatykę i strukturę wypowiedzi bez konieczności korzystania z profesjonalnej korekty. Badania jakościowe z 2025 roku przeprowadzone wśród naukowców piszących po angielsku jako w języku dodatkowym wskazują, że AI potrafi redukować część nierówności językowych, lecz jednocześnie wprowadza napięcia związane z autentycznością głosu, zależnością technologiczną i tożsamością autora. Analiza ponad miliona abstraktów z nauk społecznych z lat 2012–2024 wykazała z kolei zmniejszanie się różnic w jakości językowej między regionami. Autorzy wiążą tę konwergencję z szerszym dostępem do infrastruktury cyfrowej i narzędzi AI, choć przyjęta metoda nie pozwala przypisać tej zmiany wyłącznie jednemu czynnikowi.

Stanowi to istotny kontrapunkt dla tezy, że generatywne pisanie jest jedynie zagrożeniem. Jeśli kryterium przyjęcia artykułu nieformalnie obejmowało biegłość w akademickiej angielszczyźnie, system może usuwać barierę, która nie stanowiła o istocie kompetencji naukowej. Biolog powinien być oceniany przede wszystkim przez pryzmat biologii, a nie za przewagę wynikającą z faktu, że urodził się w środowisku anglojęzycznym. Problem pojawia się jednak wtedy, gdy korekta języka zaczyna korektę stanowiska. Badanie rozpraw doktorskich z 2025 roku sugeruje, że upowszechnienie generatywnej AI może wiązać się ze zmianą sposobu konstruowania stance: mniejszą liczbą sygnałów niepewności, większą stanowczością oraz przesunięciem od wyraźnego „głosu autora” w stronę bezosobowego „głosu tekstu”.

Autor badania podkreśla ograniczenia generalizacji, gdyż analizował konkretne subdyscypliny lingwistyki stosowanej, niemniej rezultat pozostaje teoretycznie istotny. Narzędzie nie musi fałszować faktów, aby zmienić epistemologię tekstu. Może wpływać na to, jak często badacz używa sformułowań takich jak „prawdopodobnie”, „wyniki sugerują” czy „nie możemy wykluczyć” – a zatem na retorykę niepewności, będącą kluczowym elementem naukowej ostrożności. Prowadzi to do problemu głębszego niż plagiat: homogenizacji poznawczej akademickiego głosu. Doktorant kształtuje się jako badacz nie tylko poprzez gromadzenie wyników, lecz ucząc się, kiedy można twierdzić, kiedy należy zastrzec i gdzie kończy się dowód, a zaczyna interpretacja. Jeśli system systematycznie „ulepsza” jego tekst według statystycznie dominującego wzorca akademickiej prozy,

może niepostrzeżenie ingerować w sposób, w jaki młody badacz pozycjonuje siebie wobec wiedzy. Narzędzie staje się wówczas nie korektorem, lecz cichym nauczycielem epistemicznego stylu.

Jeszcze poważniej problem wygląda na poziomie doktoratu. Rozprawa doktorska tradycyjnie pełni nie tylko funkcję publikacji wyników, lecz rytuału kwalifikacyjnego. Ma dowodzić, że kandydat potrafi samodzielnie zdefiniować problem, opanować literaturę, zastosować metodę, wytworzyć lub przeanalizować dane, bronić interpretacji i uczestniczyć w dyskursie dyscypliny. Jeśli znaczna część procesu językowego i argumentacyjnego jest wspomagana przez AI, pytanie „czy tekst został napisany przez doktoranta?" staje się zbyt prymitywne. Należy pytać, czy doktorant zachował funkcje konstytuujące samodzielność badawczą.

Uniwersytet posiada historyczny precedens takiego rozumowania. Promotor, statystyk, korektor językowy, bibliotekarz, współautor artykułu czy laborant zawsze mogli uczestniczyć w powstawaniu pracy naukowej. Akademia nie wymagała od jednej osoby wykonania wszystkich czynności; wypracowała za to normy przypisywania wkładu i odpowiedzialności. Generatywna AI wymusza rozszerzenie tej logiki. Nie chodzi o antropomorficzne wpisywanie modelu jako współautora, lecz o dokładniejsze rozliczenie ludzkiej pracy: co kandydat zaprojektował, co zweryfikował, co przeanalizował i jakie decyzje pozostały jego własne.

Równolegle rozwija się inny problem: naukowa „fabryka papieru". Materiał źródłowy przypomina, że patologie autorstwa akademickiego poprzedzają generatywną AI i obejmują ghostwriting oraz sprzedaż współautorstwa. Pozwala to uniknąć technologicznego determinizmu – AI nie stworzyła bodźców prowadzących do publikacyjnego oszustwa.

System publish or perish, mierzenie dorobku liczbą prac, presja awansowa i finansowanie zależne od wskaźników istniały wcześniej. Generatywna technologia może jednak dramatycznie obniżyć koszt produkowania tekstów przypominających naukę. Jeśli instytucja nagradza liczbę publikacji, technologia optymalizująca produkcję języka o naukowym sznycie może działać jak wzmacniacz patologicznie zaprojektowanego bodźca. Tu ponownie pojawia się prawo Goodharta.

Jeśli publikacja staje się wskaźnikiem produktywności naukowej, system zaczyna maksymalizować liczbę publikacji. Jeżeli liczba cytowań staje się miarą wpływu, powstają strategie ich maksymalizacji. Jeśli esej jest wskaźnikiem wiedzy studenta, pojawia się technologia zdolna maksymalizować jego jakość bez proporcjonalnego wzrostu wiedzy.

Przesunięcie oceny z produktu na proces myślowy

Generatywna AI nie jest więc jedynie wyzwaniem dla uczciwości jednostki, lecz testem jakości instytucjonalnych mierników, które przez dekady uznawaliśmy za wystarczające zastępniki procesów trudnych do bezpośredniej obserwacji. Odpowiedzią nie może być zatem ani całkowity zakaz, ani bezwarunkowa integracja.

Zakaz jest uzasadniony tam, gdzie samodzielne wykonanie czynności stanowi cel nauczania – podobnie jak student medycyny musi kiedyś samodzielnie przeprowadzić procedurę, mimo że w przyszłej pracy będzie korzystał z zaawansowanej aparatury. Integracja natomiast znajduje uzasadnienie tam, gdzie kompetencją docelową staje się właśnie umiejętność działania w środowisku narzędziowym. To samo ćwiczenie może więc podlegać różnym regułom na różnych etapach edukacji: początkujący powinien nauczyć się budować argumentację bez generatora, podczas gdy student zaawansowany może być oceniany z umiejętności krytycznego zarządzania generacją. Tak rozumiana edukacja wymaga od uniwersytetu radykalnej uczciwości wobec własnych celów.

Jeśli chcemy uczyć pisania, musimy czasem oceniać ten proces bez maszynowego zastępstwa. Jeśli chcemy uczyć prowadzenia badań, musimy weryfikować umiejętność sprawdzania źródeł i argumentów także w świecie, w którym AI jest powszechnie dostępna. Z drugiej strony, przygotowując absolwenta do współczesnego zawodu, absurdalne byłoby udawanie, że nigdy nie skorzysta on z modeli językowych. To nie technologia powinna decydować o zasadach zadania, lecz learning outcome. W takim modelu deklaracja użycia AI przestaje być traktowana jako rytualne samooskarżenie, a staje się informacją o metodzie pracy. Student może wskazać, że system pomógł mu poprawić język, wygenerować kontrargumenty lub skontrolować strukturę – podobnie jak badacz opisuje użyte oprogramowanie analityczne. Sama deklaracja nie rozwiązuje oczywiście problemu; szkockie badania wykazały sceptycyzm nauczycieli wobec jej skuteczności jako samodzielnego zabezpieczenia. Jej wartość rośnie dopiero wtedy, gdy towarzyszy jej możliwość sprawdzenia, czy student rozumie i potrafi obronić rezultat.

W końcu prawdziwym wyzwaniem dla uniwersytetu nie jest to, że maszyna potrafi napisać esej – uczelnia powinna być w stanie przetrwać technologię sprawnie produkującą tekst. Problem polega na tym, że przez dziesięciolecia przyzwyczailiśmy się traktować tekst jako wygodny substytut niewidzialnego procesu myślenia. Generatywna AI obnażyła słabość tego założenia.

Teraz akademia musi ponownie powiązać ocenę z rozumowaniem, źródłami, decyzjami i osobistą odpowiedzialnością autora. Nie oznacza to końca eseju, lecz może być początkiem jego odrodzenia jako formy bardziej świadomej. Esej przestaje być wyłącznie artefaktem pozostawionym na

platformie uczelnianej, a staje się częścią udokumentowanego procesu: autor formułuje problem, tworzy wersje, wchodzi w interakcję z narzędziami, odrzuca błędne propozycje, broni swoich wyborów i ponosi odpowiedzialność za tekst końcowy. Wtedy AI nie niszczy automatycznie edukacyjnego sensu pisania.

Zmusza nas jedynie do wskazania, gdzie ten sens rzeczywiście się znajdował. A gdy uniwersytet przestanie utożsamiać samą produkcję znaków z autorstwem, pojawi się problem, którego nie da się już rozwiązać wyłącznie pedagogicznie. Jeżeli tekst powstaje dzięki modelowi, promptowi, selekcji, redakcji i ludzkiej decyzji, to kto posiada prawa do rezultatu? Czy wystarcza sam pomysł? Czy prompt jest aktem twórczym? Jak wiele człowiek musi zmienić lub ukształtować, aby prawo uznało go za autora? Kryzys autorstwa przeniesie się zatem z sali seminaryjnej do urzędu praw autorskich i sądu.

> Być może największym zagrożeniem nie jest moment, w którym maszyna przejmie od nas pióro, lecz ten, w którym przestaniemy zauważać różnicę między tekstem napisanym a tekstem jedynie wygenerowanym. Czy w świecie doskonałych symulacji będziemy jeszcze potrafili docenić piękno pęknięcia i błędu, które są jedynymi szczerymi śladami ludzkiej obecności? Pytanie o autorstwo przestaje być zatem sporem technicznym, a staje się pytaniem o to, czy wciąż chcemy ponosić ciężar bycia podmiotem we własnych słowach.

Inspiracje

[1] 

"Who Wrote This" Naomi S Baron

ISBN: 9781503643574

Naomi S. Baron (ur. 27 września 1946 r.) jest wybitną amerykańską językoznawczynią i emerytowaną profesor językoznawstwa na Uniwersytecie Amerykańskim w Waszyngtonie. Uzyskała tytuł doktora na Uniwersytecie Stanforda i piastowała prestiżowe stanowiska, w tym stypendystkę Fundacji Guggenheima, stypendystkę Fulbrighta i profesor wizytującą w Centrum Studiów Zaawansowanych Nauk Behawioralnych Stanforda. Jej kariera akademicka koncentruje się na styku języka i technologii, z szeroko zakrojonymi badaniami nad komunikacją za pośrednictwem komputera, historią języka angielskiego oraz poznawczym wpływem mediów cyfrowych i drukowanych na czytanie i pisanie. Baron, płodna autorka, opublikowała wiele książek, w których bada, jak narzędzia cyfrowe i sztuczna inteligencja przekształcają ludzkie interakcje językowe, umiejętność czytania i pisania oraz myślenie krytyczne. Jej prace są powszechnie uznawane za krytyczną analizę wpływu technologii zorientowanych na wydajność na ludzką kreatywność i fundamentalną naturę autorstwa.

Naomi S. Baron (born September 27, 1946) is a prominent American linguist and Professor Emerita of Linguistics at American University in Washington, D.C. She earned her Ph.D. from Stanford University and has held prestigious positions, including Guggenheim Fellow, Fulbright Fellow, and Visiting Scholar at the Stanford Center for Advanced Study in the Behavioral Sciences. Her academic career has focused on the intersection of language and technology, with extensive research into computer-mediated communication, the history of English, and the cognitive impacts of digital versus print media on reading and writing. A prolific author, Baron has published numerous books exploring how digital tools and artificial intelligence reshape human linguistic interactions, literacy, and critical thinking. Her work is widely recognized for its critical examination of how efficiency-driven technologies influence human creativity and the fundamental nature of authorship.

American University

Literatura uzupełniająca

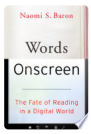
Książki

Autorzy przywoływani w artykule:

[1] "How We Read Now"

Naomi S. Baron

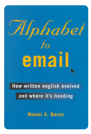
2021 | Oxford University Press | ISBN: 9780190084103



[2] "Words Onscreen"

Naomi S. Baron

2015 | Oxford University Press, USA | ISBN: 9780199315765



[3] "Alphabet to Email"

Naomi S. Baron

2000 | Psychology Press | ISBN: 9780415186858

[4] "Reader Bot"

Naomi S. Baron

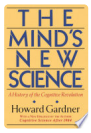
2026 | ISBN: 9781503643949

[5] "Speech, Writing, and Sign"

Naomi S. Baron

1981

Literatura pokrewna (Google Scholar):



[6] "The Mind's I"

Douglas Hofstadter

Hachette UK | ISBN: 9780786725144

[7] "Trying to Muse Rationally About the Singularity Scenario"

Douglas Hofstadter

[8] "Gödel, Escher, Bach"

Douglas Hofstadter

ISBN: 9780140179972

[9] "Creativity: Flow and the Psychology of Discovery and Invention"

Mihaly Csikszentmihalyi

Harper Collins | ISBN: 9780061844034

[10] **"Humanizing Automation"**

Ben Shneiderman

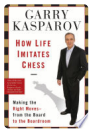
[11] **"Encounters with HCI Pioneers: A Personal History and Photo Journal"**

Ben Shneiderman

Springer Nature | ISBN: 9783031022241

[12] **"Interacting with eHealth: towards grand challenges for HCI"**

Ben Shneiderman



[13] **"How Life Imitates Chess"**

Garry Kasparov

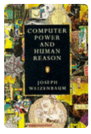
2010 | Bloomsbury Publishing USA | ISBN: 9781596918276

[14] **"SLIP"**

Joseph Weizenbaum

[15] **"ELIZA"**

Joseph Weizenbaum



[16] **"Computer Power and Human Reason: From Judgment to Calculation"**

Joseph Weizenbaum

ISBN: 9780140179118

[17] **"Data Fusion and Feature Selection for Alzheimer's Diagnosis"**

Blake Lemoine

[18] **"The Saturday Review of Politics, Literature, Science and Art"**

1866

[19] **"Academy, with which are Incorporated Literature and the English Review"**

1886

[20] **"The Saturday Review of Politics, Literature, Science, Art, and Finance"**

1866

Artykuły naukowe

Autorzy przywoływani:

[1] "Samuel Beckett's "Ohio Impromptu, Quad," and "What Where:" How It Is in the Matrix of Text and Television"

Elizabeth Klaver, Samuel Beckett

1991 | Contemporary Literature | DOI: 10.2307/1208562

Google Scholar

[2] "Creativity and artificial intelligence"

Margaret A. Boden

1998 | Artificial Intelligence | DOI: 10.1016/s0004-3702(98)00055-1

Google Scholar

[3] "Computer Models of Creativity"

Margaret A. Boden

2009 | AI Magazine | DOI: 10.1609/aimag.v30i3.2254

Google Scholar

[4] "Dimensions of Creativity"

David J. Hargreaves, Margaret A. Boden

1996 | Journal of Aesthetic Education | DOI: 10.2307/3333241

Google Scholar

[5] "WHAT IS CREATIVITY?"

Margaret A. Boden

2005 | DOI: 10.4324/9780203978627-9

Google Scholar

[6] "Ch01. Beautiful Risks"

Ronald A. Beghetto

2018 | Rowman & Littlefield Publishers eBooks | DOI: 10.5771/9781475834741-1

Google Scholar

[7] "Ch04. The Beautiful Risk of Going Off Script"

Ronald A. Beghetto

2018 | Rowman & Littlefield Publishers eBooks | DOI: 10.5771/9781475834741-35

Google Scholar

[8] **"Beyond Big and Little: The Four C Model of Creativity"**

James C. Kaufman, Ronald A. Beghetto

2009 | Review of General Psychology | DOI: 10.1037/a0013688

Google Scholar

[9] **"The Four-C Model of Creativity: Culture and Context"**

Max Helfand, James C. Kaufman, Ronald A. Beghetto

2016 | The Palgrave Handbook of Creativity and Culture Research | DOI: 10.1057/978-1-137-46344-9_2

Google Scholar

[10] **"Metafont, Metamathematics, and Metaphysics: Comments on Donald Knuth's Article "The Concept of a Meta-Font" [Reprint]"**

Douglas R. Hofstadter

2026 | Visible Language | DOI: 10.34314/2at15207

Google Scholar

[11] **"Creativity: Flow and the Psychology of Discovery and Invention"**

Mihály Csíkszentmihályi

1996

Google Scholar

[12] **"Validity and Reliability of the Experience-Sampling Method"**

Mihály Csíkszentmihályi, Reed Larson

1987 | The Journal of Nervous and Mental Disease | DOI: 10.1097/00005053-198709000-00004

Google Scholar

[13] **"The Concept of Flow"**

Jeanne Nakamura, Mihály Csíkszentmihályi

2014 | DOI: 10.1007/978-94-017-9088-8_16

Google Scholar

[14] **"Robots at work: People prefer—and forgive—service robots with perceived feelings."**

Kai Chi Yam, Yochanan Bigman, Pok Man Tang, Remus Ilieş, David De Cremer

2020 | Journal of Applied Psychology | DOI: 10.1037/apl0000834

Google Scholar

[15] "Human-machine collaboration in managerial decision making"

Tessa Haesevoets, David De Cremer, Kim Dierckx, Alain Van Hiel

2021 | Computers in Human Behavior | DOI: 10.1016/j.chb.2021.106730

Google Scholar

[16] "Establishing the importance of co-creation and self-efficacy in creative collaboration with artificial intelligence"

Jack McGuire, David De Cremer, Tim Van de Cruys

2024 | Scientific Reports | DOI: 10.1038/s41598-024-69423-2

Google Scholar

[17] "Mining for meaning: the extraction of lexico-semantic knowledge from text"

Tim Van de Cruys

2010

Google Scholar

[18] "Automatic Poetry Generation from Prosaic Text"

Tim Van de Cruys

2020 | Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics | DOI: 10.18653/v1/2020.acl-main.223

Google Scholar

[19] "AI en de taalvaardigheid van de machine"

Tim Van de Cruys

2023 | Denken over onze oorsprong | DOI: 10.2307/jj.2711707.11

Google Scholar

[20] "Section 3. INCLUDING FIGURATIVE LANGUAGE"

Arlene F. Marks, Bette J. Walker

2015 | Setting and Description | DOI: 10.5771/9781475818437-99

Google Scholar

[21] "From the Editors"

Marian Moffett, Lawrence Wodehouse

1989 | Arris | DOI: 10.1353/arr.1989.0003

Google Scholar

[22] ""Metro maps of information" by Dafna Shahaf, Carlos Guestrin and Eric Horvitz, with Ching-man Au Yeung as coordinator"

Dafna Shahaf, Carlos Guestrin, Eric Horvitz

2013 | ACM SIGWEB Newsletter | DOI: 10.1145/2451836.2451840

Google Scholar

[23] "Towards mixed-initiative access control"

Prasun Dewan, Jonathan Grudin, Eric Horvitz

2007 | 2007 International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2007) | DOI: 10.1109/colcom.2007.4553811

Google Scholar

[24] "Benchmarking Spatial Relationships in Text-to-Image Generation"

Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, E. Horvitz

2022 | arXiv.org | DOI: 10.48550/arxiv.2212.10015

Google Scholar

Artykuły pokrewne (Google Scholar):

[25] "Always On"

Naomi S. Baron

2008 | Oxford University Press eBooks | DOI: 10.1093/acprof:oso/9780195313055.001.0001

Google Scholar

[26] "Letters by phone or speech by other means: the linguistics of email"

Naomi S. Baron

1998 | Language & Communication | DOI: 10.1016/s0271-5309(98)00005-6

Google Scholar

[27] "See you Online"

Naomi S. Baron

2004 | Journal of Language and Social Psychology | DOI: 10.1177/0261927x04269585

Google Scholar

[28] "Words Onscreen: The Fate of Reading in a Digital World"

Naomi S. Baron

2015

Google Scholar

[29] "Text Messaging and IM"

Rich Ling, Naomi S. Baron

2007 | Journal of Language and Social Psychology | DOI: 10.1177/0261927x06303480

Google Scholar

[30] "Media Multitasking and Cognitive, Psychological, Neural, and Learning Differences"

Melina R. Uncapher, Lin Lin, Larry D. Rosen, Heather L. Kirkorian, Naomi S. Baron

2017 | PEDIATRICS | DOI: 10.1542/peds.2016-1758d

Google Scholar

[31] "Cross-cultural patterns in mobile-phone use: public space and reachability in Sweden, the USA and Japan"

Naomi S. Baron, Ylva Hård af Segerstad

2010 | New Media & Society | DOI: 10.1177/1461444809355111

Google Scholar

[32] "The persistence of print among university students: An exploratory study"

Naomi S. Baron, Rachelle M. Calixte, Mazneen Havewala

2016 | Telematics and Informatics | DOI: 10.1016/j.tele.2016.11.008

Google Scholar

[33] "Alphabet to Email"

Naomi S. Baron

2002 | DOI: 10.4324/9780203194317

Google Scholar

[34] "Discourse structures in Instant Messaging: The case of utterance breaks"

Naomi S. Baron

2010 | Indiana Magazine of History (Indiana University)

Google Scholar

Treść tworzy Fundacja Dobre Państwo.
Redaguje i publikuje APA ONE, autorski system redakcyjny Fundacji oparty na AI.
APA ONE Publishing System

Article ID: 7dc12209-7483-4c72-b865-b3871c40c8a9