Scripta Neophilologica Posnaniensia. Tom XXIII, strony: 121-132
Wydział Neofilologii, Uniwersytet im. Adama Mickiewicza w Poznaniu, 2023
DOI 10.14746/snp.2023.23.09

WYKORZYSTANIE SZTUCZNEJ INTELIGENCJI W OCENIE PRAC PISEMNYCH: IDEA, STAN AKTUALNY, RYZYKA, PERSPEKTYWY

Adam Pietrzykowski

Uniwersytet im. Adama Mickiewicza w Poznaniu

ORCID: 0000-0002-8353-757X

Abstrakt: Pisanie esejów odgrywa kluczową rolę w rozwoju kompetencji niezbędnych na różnych poziomach edukacji, zwłaszcza w naukach humanistycznych. Jednocześnie eseje stanowią nieodzowny element oceny dydaktycznej. Współczesne technologie przetwarzania języka naturalnego (NLP) napędzane sztuczną inteligencją (SI) oferują nauczycielom możliwość wsparcia, a nawet automatycznego oceniania prac pisemnych (Automated Essay Scoring). Ze względu na istotne implikacje pedagogiczne i społeczno-kulturowe, technologie te są dziś przedmiotem licznych analiz. W artykule przedstawiono kompleksowy przegląd technologii AES, obejmujący jej historię i pierwsze koncepcje, aktualny stan użycia, modele działania oraz kluczowe obszary krytyki. Ponadto nakreślone zostało humanistyczne podejście do włączania AES w praktykę pedagogiczną, które skupia się na optymalizacji korzyści płynących z jej wykorzystania oraz na ograniczaniu zagrożeń, jakie może ona stwarzać dla humanistycznych wartości pedagogiki.

Abstract: Writing essays plays a crucial role in developing essential competencies at various levels of education, especially in the humanities. At the same time, essays constitute an indispensable element of educational assessment. Modern Natural Language Processing (NLP) technologies powered by Artificial Intelligence (AI) offer teachers the opportunity for support and even automated grading of written assignments (Automated Essay Scoring). Due to significant pedagogical and socio-cultural implications, these technologies are the subject of numerous analyses today. The article provides a comprehensive overview of AES (Automated Essay Scoring) technologies, encompassing their history and initial concepts, current usage, operation models, and key areas of critique. Additionally, a humanistic approach to integrating AES into pedagogical practice is outlined, focusing on optimizing the benefits derived from its utilization and mitigating the threats it may pose to humanistic values of pedagogy.

Słowa kluczowe: sztuczna inteligencja, przetwarzanie języka naturalnego, automatyczna ocena esejów, AES, historia AES, krytyka AES, humanistyczne wartości pedagogiki, humanistyczne podejście wdrożeniowe

Key words: artificial intelligence, natural language processing, automated essay scoring, AES, history of AES, critique of AES, humanistic values of pedagogy, humanistic implementation approach

1. Wprowadzenie

Technologie przetwarzania języka naturalnego (*Natural Language Processing*) są dostępne w wielu obszarach życia społecznego od końca ubiegłego wieku. Jednak to osiągnięcia ostatnich lat, oparte na nowych metodach sztucznej inteligencji (AI), spowodowały prawdziwą eksplozję zainteresowania, zarówno wśród naukowców, jak i opinii publicznej. Bezpłatne udostępnienie tzw. dużych modeli językowych (LLM), które wykorzystują głębokie sieci neuronowe i ogromne zbiory danych, takie jak ChatGPT, Bard czy StableLM, ożywiło nadzieje technoutopistów na świat wolny od przymusu pracy, a jednocześnie obudziło lęki technologicznych katastrofistów, którzy widzą w nich zagrożenie dla społecznego ładu i nadchodzący zmierzch ludzkości. Z jednej strony tę dualną optykę tłumaczą obecne w kulturze apokaliptyczne i utopijne wizje dotyczące roli techniki, które – jak wskazuje Odo Marquard – są efektem znacznego przyśpieszenia rozwoju techniki względem zachowawczej i „powolnej" kultury (Marquard, 1994:90). Z drugiej, realna ocena wpływu, jaki technologie te wywierają na życie społeczne mierzona zmianami na rynku pracy, sposobie wykonywania określonych profesji i funkcjonowania całych obszarów życia społecznego. Nie bez powodu tego rodzaju technologie określa się mianem przełomowych innowacji (*disruptive technologies*) (zob. Christensen, 1997), co w dosłownym tłumaczeniu oznacza „technologie zakłócające" – w domyśle – istniejące praktyki społeczne i wzory kulturowe.

Edukacja jest jednym z obszarów, w których technologie NLP zyskują coraz większe znaczenie. Obecnie dyskusje wokół sztucznej inteligencji skupiają się na dużych modelach językowych, takich jak wspomniany ChatGPT, który uważany jest za „złoty standard" w generowaniu treści. Niemniej na uwagę zasługują również zastosowania sztucznej inteligencji, które znajdują się na przeciwległym biegunie, mianowicie systemów służących do ewaluacji i oceny wypowiedzi pisemnych. Tym bardziej, że pomimo ich niezwykle dynamicznego rozwoju w ostatniej dekadzie, są one niemal nieobecne w polskim dyskursie akademickim.

2. NLP w ewaluacji tekstu – historia i stan aktualny

Narzędzia przetwarzania języka naturalnego (NLP) służące do oceny i ewaluacji tekstu są obecne w edukacji już od końca lat 90. W literaturze przedmiotu można zauważyć ich podział na dwie kategorie: pierwsza obejmuje narzędzia do automatycznego oceniania, które określane są terminami takimi jak: Automated Essay Scoring (AES), Automated Essay Grading (AEG) oraz Automatic Short-Answer Scoring (ASAS) (zob. Bai i Stede, 2022; Woods et al., 2017). Drugą kategorią są narzędzia ewaluujące tekst występujące pod szyldem Automatic Essay Evaluation (AEE) i Automated Writing Evaluation (AWE). Narzędzia z pierwszej

kategorii mają na celu głównie dostarczanie dodatkowej oceny, szczególnie przydatnej w przypadku egzaminów krajowych, które zawierają wypowiedzi pisemne, takie jak krótkie i długie odpowiedzi na pytania testowe oraz eseje. Ponadto, narzędzia te są używane do zastąpienia człowieka w sytuacjach, gdy nauczyciel z założenia jest nieobecny, na przykład w masowej edukacji online. Natomiast narzędzia ewaluujące tekst, które oferują natychmiastową informację zwrotną, zostały stworzone w celu wspierania uczniów i studentów podczas tworzenia prac pisemnych oraz dostarczania nauczycielom dodatkowych metryk dotyczących ich jakości (Neslihan et al., 2020). Warto zauważyć, że rozwiązania AES i AEE stanowią często dwa moduły jednego systemu, co sprawia, że powyższy podział jest czysto teoretyczny.

Szczególnie obiecującym obszarem zastosowania rozwiązań AES i AEE jest edukacja online. Zwłaszcza w modelach asynchronicznych opartych o uczenie się we własny tempie (*self-paced*), w tym w masowych kursach typu MOOC (*Massive Open Online Courses*), które eliminują obecność nauczyciela w procesie kształcenia. Ponadto, są one wykorzystywane także w pełnowartościowym e-learningu akademickim m.in. do ewaluacji wypowiedzi na forum dyskusyjnym (Rutner i Scott, 2022).

Idea AES pojawiła się po raz pojawiła się w latach 60. w pracach Ellisa Page'a, profesora psychologii edukacyjnej. Page, który podczas studiów pracował jako nauczyciel angielskiego dostrzegł, iż możliwości oferowane ze strony maszyn obliczeniowych i rodzącej się lingwistyki komputerowej mogą wesprzeć nauczyciela w żmudnym procesie sprawiania prac zaliczeniowych (Potts, 2005). Jednak dla Page'a zastosowanie komputera oznaczało coś więcej niż tylko odciążenie. To także większa zgodność oceny między nauczycielem, a modelowanym na nim komputerowym sędzią (Page, 1966). W efekcie miało to gwarantować bardziej obiektywny wynik i konsensus sprawdzających, co jest trudniejsze do uzyskania między dwojgiem oceniających nauczycieli. W swoich badaniach wykorzystywał ręcznie napisane eseje, które następnie przerabiał na karty perforowane i wprowadzał do jednej z pierwszych, maszyn obliczeniowych zajmujących kilka pomieszczeń. Program, który Page stworzył w 1966 analizował eseje pod kątem występowania określonych cech (*features*), które według założeń jego zespołu, oznaczać miały dobre pisarstwo. Była to m.in. ilość długich zdań oraz różnorodność językową, które wskazywać miały na szeroki zasób słów i umiejętność wyrażania złożonych koncepcji. W 1968, po dwóch latach badań, opublikował wyniki badań nad pierwszym programem komputerowy Project Essay Grade (PEG) służącym temu zadaniu (Page, 1968). PEG nie był nieefektywny z uwagi na ogromne zasoby jakich wymagał do działania, jednak dowodził, iż sama idea komputerowej analizy tekstu jest możliwa do przeprowadzenia, a ocena komputera i człowieka okazują się zbliżone. Dopiero w drugiej połowie lat 90. wzrost mocy obliczeniowe komputerów oraz rozwój metod nauczania maszynowego i algorytmiki przyniosły pierwsze komercyjne realizacje idei automatycznej oceny esejów (AES).

Jak się okazuje automatyczne rozwiązania oceniające wypowiedzi pisemne bynajmniej nie znajdują się na marginesie praktyki edukacyjnej. W Stanach Zjednoczonych są obecnie masowo wykorzystywane w obszarze edukacji podstawowej i średniej. W 2019 roku systemy AES obecne były w 21 stanach USA, z czego tylko w trzech prace były ponownie sprawdzane przez człowieka (Feathers, 2019). Zdecydowanie rzadziej, głównie w formie wspomagających asystentów, stosowane są w Niemczech, Chinach, Japonii i Meksyku (Bai i Stede, 2022).

Wśród dziesiątek systemów AES kilkanaście ma charakter komercyjny i jest wykorzystywana w praktyce. Jednym z nich jest oprogramowanie Intelligent Essay Assessor brytyjskiej firmy Pearsons. Do 2018 zdążyło ocenić aż 34 mln studenckich esejów w krajowych testach kwalifikacyjnych wysokiej rangi (Smith, 2018). Innym systemem wykorzystywanym na szeroką skalę w USA jest e-rater organizacji non-profit Educational Testing Service (ETS), która stosuje je w testach SAT, GRE i TOEFL. Także współczesna wersja Project Essay Grade (PEG), obecnie należąca do firmy Measurement Incorporated, jest wykorzystywana by rocznie ocenić blisko 10 mln prac uczniów i studentów w USA (Measurement Inc., n.d.). W przypadku masowej edukacji przez Internet powołana przez Harvard University i MIT platforma edX, oferuje narzędzie Discern, które ocenia prace tysięcy studentów zapisanych na darmowe kursy online (Markoff, 2013).

3. Modele i skuteczność działania systemów AES

Modus operandi systemów automatycznej oceny esejów opiera się na trzech filarach: 1) zbiorze cech tekstu (*features*), które podlegają pomiarowi, 2) zbiorze danych składających się z ocenionych przez człowieka wypowiedzi pisemnych służący do wytrenowania systemu, 3) technikach modelowania matematycznego (Ramesh i Sanampudi, 2022).

Pierwsze systemy AES i wzorowane na nich rozwiązania wprowadzały swoistą inżynierię cech tekstu. Programy takie jak PEG Ellisa Page'a czy współczesny e-rater firmy ETS analizują tekst pod kątem wskazanych przez ekspertów cech, takich jak poprawność gramatyczna, długości zdań czy bogactwo leksykalne, uznając je za korelaty jakości pracy (Woods et. al., 2017). Analizowane przez programy cechy eksperckie umieścić można w trzech kategoriach: statystyczne, związane ze stylem oraz związane z treścią. Ramesh i Sanampudi (2022) opracowali listę cech w ramach wymienionych kategorii, które są stosowane w większości systemów AES (Tabela 1).

Poszukując wyższej skuteczności w prognozowaniu ocen wystawionych przez człowieka kolejne systemy wykorzystywały najnowsze rozwiązania algorytmiczne. Niezwykle skuteczne okazały się modele oparte o głębokie sieci neuronowe (*deep neural networks*) i uczenie głębokie (*deep learing*), które definiują własne, nisko poziomowe (*low-level*) syntaktyczne i semantyczne cechy tekstu (Ramesh i Sanampudi, 2022). Stosują one techniki modelowania wielowymiarowego

Statystyczne cechy tekstu	Cechy analizujące styl	Cechy analizujące treść
Długość eseju – liczby słów	Struktura zdań	Spójność między zdaniami w dokumencie
Długość eseju – liczby zdań	Części mowy	Nakładanie się (odpowiedzi i wzorca odpowiedzi)
Średnia długość zdania	Interpunkcja	Znaczenie informacji
Średnia długość wyrazu	Gramatyka	Semantyczna rola wyrazów
N-gram	Operatory logiczne	Poprawność
	Słownictwo	Spójność
		Zdania wyrażające kluczowe koncepcje

Tabela 1. Rodzaje cech wypowiedzi pisemnej analizowane przez programy AES w modelach eksperckich. Za Ramesh i Sanampudi (2022).

przez co odchodzą od podejścia, zgodnie z którym cechy modelu powinny naśladować ludzkie rozumowanie (Woods et. al, 2017). Skuteczne okazują się także modele wytrenowane wcześniej (*pre-trained*) jak np. BERT firmy Google i jego pochodne, zaś za najlepsze uważa się dziś modele hybrydowe łączące ekspercką inżynierię cech z głębokimi sieciami neuronowymi (Li et al., 2023).

Aby system AES mógł pełnić swoją rolę, a więc dostarczać oceny zbliżonej do oceny człowieka, konieczne jest uczenie go na dużym zbiorze danych. Według firmy Pearsons do wytrenowania systemu oceniającego jedno otwarte pytanie potrzebne jest co najmniej 2000 do 5000 sprawdzonych prac (Barshay, 2022). Współcześnie istnieje co najmniej kilkanaście anglojęzycznych zbiorów danych, z których korzysta się do trenowania modeli AES (Ramesh i Sanampudi, 2022). Najczęściej wykorzystywany jest zbiór danych anglojęzycznych udostępniony w 2012 roku przez platformę Kaggle, spółkę-córkę Google, w ramach konkursu Automatic Student Assessment Prize (ASAP). Ramesha i Sanampudi (2022) szacują, że dane ASAP są dziś wykorzystywane do uczenia 90% systemów działających w języku angielskim. Zbiór składa się zarówno z danych do trenowania systemów oceniające eseje (ASAP-AES), jak i krótkich odpowiedzi pisemnych (ASAP-SAS). W skład pierwszego wchodzi blisko 13 000 esejów o charakterze narracyjnym, argumentacyjnym oraz bazującym na źródłach, napisanych przez amerykańskich uczniów z klas 7-10 (Mathiasa i Bhattacharyi, 2020).

Ostatnim elementem jest modelowanie matematyczne. Systemy analizujące cechy statystyczne wykorzystują regresję liniowa, zaś w przypadku cech związanych ze stylem czy treścią stosowane są modele sieci neuronowych lub utajona analiza semantyczna (*Latent Semantic Analysis*).

Próbując określić skuteczność oceny systemów AES należy zauważyć, że podobnie jak z każdą inną metodą oceniania muszą one odznaczać się trzema cechami: ważnością (*validity*), uczciwością (*fairness*) i wiarygodnością (*reliability*) (Chung et al., 2003). W przypadku systemu komputerowego kluczowa jest wiarygodność, która najczęściej oznacza zgodność oceny wystawionej przez system z tzw. prawdziwą oceną (*true score*), czyli uśrednioną oceną dwóch ludzi (*inter--rater agreement*). Do określenia zgodności stosuje się metody statystyczne, takie jak: ważona kappa, średni bezwzględny błąd procentowy oraz współczynnik korelacji Pearsona (Shehab et al., 2016). Dla systemów uznanych za wiarygodne, a tym samym skuteczne, przyjmuje się odsetek zgodności z człowiekiem na poziomie co najmniej 75% (Graham et al., 2012).

Dla badań nad skutecznością systemów AES w naśladowaniu oceny człowieka przełomowym momentem był rok 2012, gdy fundacja Hewlett zorganizowała konkurs Automated Student Assessment Prize (ASAP). W konkursie studenci z całego świata rywalizują z największymi komercyjnymi dostarczycielami rozwiązań AES o tytuł najbardziej skutecznego modelu AES i ASAS (UoA, 2012). Najlepsze systemy okazują się przewidywać nieco powyżej 80% ocen wystawionych przez człowieka (Paruchuri, 2013). Dziś eksperymentalne rozwiązania oparte o głębokie uczenie się i głębokie sieci neuronowe mogą być jeszcze skuteczniejsze osiągając zbieżność z człowiekiem na poziomie 95% (Nguyen i Dery, 2016).

4. Punkty sporne

Wysoka skuteczność współczesnych systemów AES nie oznacza, że są one pozbawione wad i problemów. Istnieje wiele punktów spornych dotyczących sposobu ich funkcjonowania. Krytyka dotyczy głównie tego, co w istocie owe systemy mierzą, czy są wiarygodne, czy potrafią przekazać wartościowy feedback oraz jak ogólnie oddziałują na edukację.

Pierwszy z obszarów krytyki dotyczy wiarygodności oceny systemu. W 2012 roku Mark Shermis przeprowadził badanie systemów AES na próbie 22 tys. krótkich esejów uczniów szkół podstawowych i średnich ocenianych niezależnie przez program komputerowy i przeszkolonych czytelników (Shermis, 2014). Badanie to wykazało brak istotnej statystycznie różnicy, co wywołało zarówno skrajnie entuzjastyczne jak i sceptyczne reakcje. Wśród sceptyków znalazł się Les Perelman, wykładowca pisarstwa i składu na MIT, który wskazał istotne błędy metodologiczne badania oraz ograniczenia podważające sens tworzenia rozwiązań automatyzujących ocenianie prac pisemnych (Perelman, 2013). Dowodził, że algorytmy nie oceniają jakości eseju pod kątem znaczenia i zgodności z faktami, co oznacza, że łatwo je zmanipulować. Dowiódł tego tworząc ze swoimi studentami generator esejów BABEL (Basic Automatic B.S. Essay Language), który był w stanie oszukać każdy komercyjny system AES tworząc nonsensowne i fikcyjne treści uzyskując przy tym najwyższe oceny.

Analiza Perelmana stała się postawą do zorganizowania petycji online, w której krytyczne stanowisko na temat wykorzystania systemów AES w kluczowych egzaminach i testach końcowych wyraziło ponad 4000 nauczycieli i wykładowców akademickich, w tym Noam Chomsky. Apel środowiska nie wpłynął jednak w żaden sposób na decydentów w USA. W 2020 roku Perelman ponownie przeanalizował powszechnie wykorzystywany w USA system e-rater dowodząc, iż nadal nie rozpoznaje on nonsensownych zdań i zwrotów, co powinno dyskwalifikować go z użycia w egzaminach pisemnych wysokiej rangi (Perelman, 2020).

Także sposób uczenia się systemów AES, a więc trenowania ich na zbiorze esejów i skojarzonych z nimi ocen, budzi kontrowersje. Badania algorytmów uczących się dowodzą, iż system wytrenowany na zbiorze ocenionych przez konkretnego człowieka prac przejmuje również jego uprzedzenia (Amorim et al., 2018). Dotyczą one najczęściej grup społecznych oraz płci, co okazuje się mieć odzwierciedlenie w wystawionych ocenach. Co więcej, jak zauważa Emily Bender, systemy uczące się maszynowo wzmacniają ów efekt (Feathers, 2019). Widać to na przykładzie narzędzia e-rater firmy ETS wykorzystywanego m.in. w egzaminach wstępnych na amerykańskie uczelnie (GRE) oraz do sprawdzania kompetencji językowych (TOEFL). Badania ETS wykazały, że w porównaniu z człowiekiem narzędzia te faworyzują osoby z Chin, zaś dyskryminują Afroamerykanów, osoby pochodzenia arabskiego oraz Latynosów (Feathers, 2019). Obiektywność systemu okazuje się zatem zależna od obiektywności osoby, na której jest on modelowany. Oznacza to, że doskonała maszyna oceniająca, która byłaby wolna od jakichkolwiek uprzedzeń nie jest możliwa. Powyższy fakt budzi nieuchronnie wątpliwości natury etycznej, zwłaszcza w przypadku stosowania tego rodzaju rozwiązań do egzaminów kwalifikacyjnych, w których uczestniczą osoby z różnych warstw społecznych i kultur.

Niedoskonałość modeli dotyczy także kwestii języka w jakim system jest trenowany. Nieliczna literatura sugeruje, że w różnych językach zgodność systemu z oceną wystawioną przez człowieka oraz jego wiarygodność rozumiana jako rozpoznawanie fikcyjnych i nonsensownych treści, może być różna. W jednym z badań nad systemami AES operującymi w języku niemieckim próbowano przewidzieć zarówno ocenę końcową, jak i punktację cząstkową studenckich esejów (Horbach et al., 2017). Niestety zastosowane przez badaczy dwa różne podejścia, model sieci neuronowych oraz nadzorowany model nauczania maszynowego oparty o cechy, okazały się znacznie mniej skuteczne od przewidywań. Autorzy badania przypisali uzyskany wynik zarówno złożoności języka niemieckiego, jak i wysokiemu poziomowi umiejętności pisania wykazanemu w esejach.

Nie rozumiejąc treści wypowiedzi pisemnej systemy AES nie mogą zaoferować tego, co każdy student może otrzymać od swojego wykładowcy – precyzyjnego feedbacku na temat pracy (Ke i Ng, 2019). Należy zauważyć, że ten aspekt systemów AES będzie się różnił się w zależności od sposobu działania jego modelu. Modele oparte o cechy wskazane przez ekspertów, określane jako „białe skrzynki" (*white-box*) ze względu na pełną wiedzę o tym jak działają, oferują informacje

zwrotną na temat statystycznych, syntaktycznych czy leksykalnych walorów pracy. W przypadku bardziej skutecznych modeli opartych o głębokie sieci neuronowe oraz modele wytrenowane wcześniej (*pre-trained*) mamy do czynienia z „czarnymi skrzynkami" (*black-box*). Oznacza to, że nie można dokładnie ustalić sposobu działania systemu, a zatem nie ma możliwości uzyskania jakiejkolwiek informacji zwrotnej o składowych oceny. Według niektórych krytyków ta sytuacja może powodować spadek motywacji do pisania, zwłaszcza wśród studentów, których prace będą oceniane przez drugi typ systemów (Dikli i Semire, 2006).

Wreszcie zapytać należy o nauczyciela, którego aktywność zawodowa podlegać będzie zmianom wywołanym przez omawiane technologie. W przypadku krótkich wypowiedzi w testach o charakterze odtwórczym systemy AES niewątpliwie pozwalają oszczędzić czas, który może być spożytkowany na inne, bardziej wartościowe zadania. Poważne wątpliwości dotyczą sytuacji, gdy system AES miałby zastąpić nauczyciela w ocenie twórczych prac, zwłaszcza własnego ucznia czy studenta. To właśnie w eseju nauczyciel sprawdza postępy swojego podopiecznego, ale także poznaje jego osobowość w sposób niedostępny na zajęciach grupowych. Pisanie jako forma aktywności jest bowiem osadzona w intymności. Udziela czytelnikowi dostęp do części swojego prywatnego świata. Dlatego dla nauczyciela esej jest narzędziem budowania relacji z uczniem, co pięknie pokazał np. nagrodzony Oskarami film Darrena Aronofskiego *The Whale*. Aktualnie obawa o eliminację tego aspektu edukacji jest znikoma i dotyczy wyłącznie modeli akademickiej edukacji online, które realizowane są w neoliberalnym paradygmacie technologicznym (MOOC). Zorientowanie na skalowalność i redukcje kosztochłonności chętnie korzystają bowiem z wszelkich technologii automatyzujących.

5. Perspektywy

Przeprowadzona analiza przybliżyła różne konteksty technologii automatycznego oceniania wypowiedzi pisemnych ukazując zarówno przestrzeń szans, jak i ryzyka. Jakie perspektywy rysują się zatem dla technologii AES? W opinii Ananta Agarwala, CEO platformy edX i profesor MIT, rozwój systemów tego typu jest nieuchronny, gdyż oferują one nie tylko oszczędność czasu nauczyciela, ale także natychmiastową informację zwrotną, która ma szczególne znaczenie w kształceniu online (Markoff, 2013). Powyższa opinia wydaje się nie dostrzegać zarysowanych wcześniej problemów związanych z AES. Wyrażona jest bowiem z punktu sektora EdTech, generującego miliardowe zyski każdego roku i lobbującego w USA za bardziej odważnymi wdrożeniami systemów SI w edukacji. Trudno uznać ją za obiektywną i mającą na uwadze przede wszystkim dobro edukacji. Dlatego warto uzupełnić ją o inne perspektywy.

Dla utalentowanego inżyniera, laureata konkursu studenckiego ASAP Vikasa Paruchuri, nauczyciele powinni być zaznajomieni z systemami AES, żeby

wiedzieć jak najlepiej je wykorzystać, ale też czy w ogóle to robić (Vikas, 2013). W jego opinii, o tym czy należy je stosować w konkretnej sytuacji, np. w przypadku otwartych pytań w teście czy też by ocenić szkic pracy zaliczeniowej, powinni decydować samodzielnie kierując się potencjalną korzyścią jaką odniesie student.

Widoczne powyżej upodmiotowienie nauczycieli w procesie społecznej recepcji technologii AES kontrastuje z idealistyczną i atechniczną optyką Erica Thomasa, wykładowcy dziennikarstwa na University of Kansas. Thomas (2023) uważa, że nauczyciele i instytucje powinny oprzeć się pokusie efektywności oferowanej przez technikę. Zamiast tego proponuje ukazać studentom wartość uczenia się przez trudne, a często pełne pokory próby napisania czegoś znaczącego, jednocześnie w pełni angażując się w ten proces.

Wyważoną i zorientowaną pedagogicznie postawę w aspekcie wykorzystania technologii AES w edukacji reprezentuje z kolei australijskie środowisko edukacyjne. W 2018 roku National Education Council, m.in. na podstawie raportów sporządzonych przez Lesa Perelmana, uchyliła plany Ministra Edukacji, które zakładały wprowadzenie systemu AES w krajowych egzaminach oceny kompetencji (NAPLAN) (Koziol et. al., 2018). W 2022 roku rozpoczęto kolejne podejście do wdrożenia AES w australijskim systemie szkolnictwa uzależniając je tym razem od wyników przeprowadzonych badań. Biała księga opracowana przez naukowców z University of Sidney sugeruje, iż AES może być wartościowe zarówno dla uczniów jak i nauczycieli pod warunkiem spełnienia określonych warunków (Gulson et al., 2022), w tym:

- Ustanowienia jednolitych, etycznych wytycznych dla szkół kupujących systemy AES od komercyjnych firm Edtech,
- Stworzenia wytycznych dotyczących najlepszych praktyk w zakresie wykorzystania sztucznej inteligencji w szkołach,
- Powołania niezależnego organu doradczego ds. oceny na dużą skalę,
- Ustanowienia ramy teoretycznej dla analizy ryzyka, za pomocą której klasyfikowane będą potencjalne szkody wynikające z przyjęcia AES w konkretnych kontekstach oceniania i działania szkoły,
- Wyciągania wniosków z wdrożeń AES na dużą skalę w USA i innych regionach, uwzględniając implikacje etyczne, prawne i finansowe oraz wpływ na proces decyzyjny w szkołach, uczelniach i regionach,
- Przed wdrożeniem wykorzystanie wiedzy interesariuszy z wielu lokalizacji i poziomów podejmowania decyzji, w tym nauczycieli,
- Niestosowania AES w testach krajowych wysokiej rangi, których wyniki mają istotne znaczenie dla poszczególnych osób lub szkoły.

Troska o dobro edukacji i zachowanie humanistycznych wartości pedagogiki, która przebija przez powyższe rekomendacje, wydaje się kierunkiem godnym naśladowania. Współbrzmi z tym, co John Naisbitt zawarł w prostej frazie

„high tech, high touch", a więc by rozwojowi techniki towarzyszyło większe wyczulenie na kwestie wartości, tak by „wspierać technologie, które zachowują nasze człowieczeństwo i odrzucać te, które je podważają" (Naisbitt, 1999:26). Jeśli zatem systemy AES nie będą nigdy w stanie dostrzec iskry geniuszu, autentyczności czy prawdy stojącej za autorem dobrego eseju, przy okazji rozwijając relacje międzyludzkie, to w czym tak naprawdę ważnym mogą nam pomóc? Z pewnością istnieją mniej znaczące i pracochłonne obszary edukacji, które odniosą z pewnością pewną korzyść. Jednak zbyt szerokie ich użycie, do czego zmierza obecnie system ewaluacji w USA, doprowadzić może do edukacyjnej dystopii: edukacji bez oceniającego i ocenianego, opartej wyłącznie na interakcji systemów sztucznej inteligencji, które w imieniu każdej ze stron będą tworzyć i sprawdzać prace.

Bibliografia

Amorim, E., Cançado, M., Veloso, A. (2018). „Automated essay scoring in the presence of biased ratings". W: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, 229-237. New Orleans: Association for Computational Linguistics.

Bai, X., Stede, M. (2022). „A survey of current machine learning approaches to student free-text evaluation for intelligent tutoring". *International Journal of Artificial Intelligence in Education* 28, 1–39. https://doi.org/10.1007/s40593-022-00323-0.

Barshay, J. (2022). *PROOF POINTS: A smarter robo-grader. The Hechinger report*,
https://hechingerreport.org/proof-points-a-smarter-robo-grader.

Christensen, C.M. (1997). *The innovator's dilemma: how new technologies cause great firms to fail*. Harvard Business School Press.

Chung, G.K.W.K., Baker, E.L. (2003). „Issues in the reliability and validity of automated scoring of constructed responses". W: M.D. Shermis & J. Burstein (red.), *Automated essay scoring: A cross-disciplinary perspective*. 23–40. Lawrence Erlbaum Associates Publishers.

Dikli, S. (2006). „An overview of automated scoring of essays". *Journal of Technology, Learning, and Assessment*, 5(1), 1-36.

Feathers, T. (2019). „Flawed Algorithms Are Grading Millions of Students' Essays*". Vice,* https://www.vice.com/en/article/pa7dj9/flawed-algorithms-are-grading-millions-of-students-essays.

Graham, M., Milanowski, A., Miller, J. (2012). „Measuring and promoting inter-rater agreement of teacher and principal performance ratings". *The Center for Educator Compensation and Reform*, 1-37.

Gulson, K., Thompson, G., Swist, T., Kitto, K., Rutkowski, L., Rutkowski, D., Hogan, A., Zhang, V., Knight, S. (2022). *Automated essay scoring in Australian schools: key issues and recommendations*. *White Paper*. University of Sydney, Australia.

Horbach, A., Ding, Y., Zesch, T. (2017). „The influence of spelling errors on content scoring performance". W: *Proceedings of the 4th workshop on natural language processing techniques for educational applications*, 45–53. Taipei: Asian Federation of Natural Language Processing.

Ke, Z., Ng, V. (2019). „Automated essay scoring: a survey of the state of the art". *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence Survey track*. 6300-6308, https://doi.org/10.24963/ijcai.2019/879.

Kolowich, S. (2012). „A win for the Robo-Readers", *Inside Higher Ed*, https://www.insidehighered.com/news/2012/04/13/large-study-shows-little-difference-between-human-and-robot-essay-graders.

Koziol, M., Singhal, P., Cook, H. (2018). „Computer says no: governments scrap plan for 'robot marking' of NAPLAN essays". *The Sydney Morning Herald*, https://www.smh.com.au/politics/federal/computer-says-no-governments-scrap-plan-for-robot-marking-of-naplan-essays-20180129-h0py6v.html.

Li, F, Xi, X, Cui, Z, Li, D, Zeng, W. (2023). „Automatic essay scoring method based on multi-scale features". *Applied Sciences* 13(11), 6775. https://doi.org/10.3390/app13116775.

Markoff, J. (2013). „Essay-grading software offers professors a break". *The New York Times*, https://www.nytimes.com/2013/04/05/science/new-test-for-computers-grading-essays-at-college-level.html?pagewanted=all&_r=0.

Marquard, O. (1994). *Apologia przypadkowości*. Warszawa: Oficyna Naukowa.

Measurement Inc. (n.d.). Pobrane z https://www.measurementinc.com/services#scoring-services.

Naisbitt, J., Naisbitt, M., Philips, D. (1999*). High tech high touch*. New York: Broadway Books.

Nguyen, H., Dery, L. (2016). „Neural networks for automated essay grading". *Report for CS224d: Deep Learning for Natural Language Processing*. https://cs224d.stanford.edu/reports/huyenn.pdf.

Page, E.B. (1966). „The imminence of... grading essays by computer". *The Phi Delta Kappan* 47(5), 238-243.

Page, E.B. (1968). „The use of the computer in analyzing student essays". *International Review of Education* 14(3), 253-263.

V. Paruchuri (2013). *On the automated scoring of essays and the lessons learned along the way*, https://www.vikas.sh/post/on-the-automated-scoring-of-essays.

Perelman, L.C. (2013). „Critique of Mark D. Shermis & Ben Hamner. Contrasting state-of-the-art automated scoring of essays: analysis". *The Journal of Writing Assessment* 6(1).

Perelman, L.C. (2020). „The BABEL Generator and E-Rater: 21st Century Writing Constructs and Automated Essay Scoring (AES)". *Journal of Writing Assessment* 13(1).

Potts, M. (2005). „Ellis Page, 81, a developer of computerized grading, dies". *The New York Times*, https://www.nytimes.com/2005/05/23/us/ellis-page-81-a-developer-of-computerized-grading-dies.html.

Shermis, M.D. (2014). „State-of-the-art automated essay scoring: competition, results, and future directions from a United States demonstration". *Assessing Writing* 20, 53-76. https://doi.org/10.1016/j.asw.2013.04.001.

Mathias, S., & Bhattacharyya, P. (2020). „Can neural networks automatically score essay traits? W: *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. 85–91. Seattle: Association for Computational Linguistics.

Ramesh, D., Sanampudi, S.K. (2022). „An automated essay scoring systems: a systematic literature review". *Artificial Intelligence Review* 55, 2495–2527. https://doi.org/10.1007/s10462-021-10068-2.

Rutner, S.M., Scott, R.A. (2022). „Use of artificial intelligence to grade student discussion boards: an exploratory study". *Information Systems Education Journal* 20(4), 4-18.

Shehab, A., Elhoseny, M., Hassanien, A.E. (2016). „A hybrid scheme for automated essay grading based on LVQ and NLP techniques. W*: 12th International Computer Engineering Conference (ICENCO)*. Kair, 65-70.

Smith, T. (2018). „More states opting to 'Robo-Grade' student essays by computer". *National Public Radio*, https://www.npr.org/2018/06/30/624373367/more-states-opting-to-robo-grade-student-essays-by-computer?t=1603620181809.

Süzen, N., Gorban, A.N., Levesley, J., & Mirkes, E.M. (2020). „Automatic short answer grading and feedback using text mining methods". *Procedia Computer Science* 169, 726-743. https://doi.org/10.1016/j.procs.2020.02.171.

Taghipour, K., & H.T. Ng (2016). „A neural approach to automated essay scoring. W*: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* 1882–1891.

Thomas, E. (2023). „Grading can overwhelm teachers as the semester ends. Does AI present a better way?". *Kansas Reflector*, https://kansasreflector.com/2023/04/14/grading-can-overwhelm-teachers-as-the-semester-ends-does-ai-present-a-better-way.

UoA. (2012). *Man and Machine: better writers, better grades*. https://cacm.acm.org/news/148670-man-and-machine-better-writers-better-grades.

Woods, B., Adamson, D., Miel, S., & Mayfield, E. (2017). „Formative essay feedback using predictive scoring models". W*: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3097983.3098160.