„Teologia i Moralność", Volumen 20 (2025), numer 3 (40), s. 271–278
ISSN 1898-2964 eISSN 2450-4602
DOI: 10.14746/TIM.2025.40.3.19

Sztuczna Inteligencja – moralna czy pozytywna?

Artificial Intelligence – moral or positive?

Damian Szczęch
Szkoła Doktorska KUL, Filozofia
Katolicki Uniwersytet Lubelski Jana Pawła II, Polska
damian.szczech@kul.pl
https://orcid.org/0000-0003-2290-1726

Jana Schaich Borg, Walter Sinnott-Armstrong, Vincent Conitzer. 2024. *Moral AI: and how we get there*, Pelican Books, ss. 236 (z przypisami i indeksem 270)

Książka składa się z Przedsłowia (*Preface*), Wprowadzenia, siedmiu rozdziałów oraz Wniosków. Na końcu umieszczono przypisy oraz indeks pojęć i nazwisk. W przedsłowiu autorzy opisali swoje zainteresowania badawcze, podkreślając że reprezentują różne dyscypliny (neuronauki, filozofię i *computer science*). Wspólny jest im optymizm związany z AI oraz troska o moralne problemy dotyczące jej rozwoju. Książka ma na celu wskazanie tych problemów i przedstawienie możliwych do realizacji sugestii (*actionablesuggestions*) dotyczących takiego prowadzenia rozwoju AI, aby minimalizować wyrządzane szkody moralne.

We wstępie autorzy zauważają powszechność strachu przed AI, która może wpływać na nasze zachowania, naruszać wolność i prywatność. Wymieniają kwestie kluczowe dla zrozumienia jej natury i ograniczeń oraz zapowiadają podjąć je w kolejnym rozdziale (czego niestety nie zrobili): *kiedy program komputerowy staje się AI? Czy AI naprawdę jest inteligencją? Czy AI może być świadoma? Czy może być kreatywna, czy tylko bezmyślnie wykonuje instrukcje? Czy może wykonywać nowe działania, czy jest ograniczona do tych samych zadań, które wcześniej wykonywali ludzie? Które rodzaje działań może wykonywać, a których nie? Czy AI może stać się bardziej inteligentna niż ludzie?* (s. XVIII–XIX). Twierdzą, że gdy ustalimy czym AI jest, a czym nie jest, koniecznym stanie się zrozumienie, jak oddziałuje z „podstawowymi moralnymi

wartościami" (s. XIX–XX), jak np.: bezpieczeństwo (utrata kontroli nad naszym światem lub jego częścią), równość (*bias* w ocenie ryzyka na podstawie rasy, zamożności i płci), prywatność (kazus Cambridge Analytica), wolność (inwigilacja utrudniająca działanie ruchom religijnym), przejrzystość (ocena wydajności w pracy, prac uczniów, zdolności kredytowej), oszustwa (*deception*; Deepfake i *fake news* wpływające na politykę) [podejrzewam, że chodziło o wolność od oszustw, a nie o oszustwo jako wartość]. AI może być budowana i używana w sposób bezpieczny i moralny, o ile wskazane rodzaje moralnych problemów zostaną wzięte pod uwagę. Niestety autorzy nie wskazali na podstawie jakiej aksjologii i etyki określili te podstawowe wartości i problemy.

Rozdział 1. pt. „Czym jest AI?" rozpoczyna stwierdzenie, że AI nie ma ogólnie akceptowanej definicji. Autorzy przyjęli taką: „System komputerowy, który dla danego zestawu celów, określonych przez człowieka może dokonywać prognoz, zaleceń lub decyzji wpływających na rzeczywiste, lub wirtualne środowiska z wystarczającym poziomem niezawodności (*reliability*)" (s. 2). Wprowadzili też pojęcia: wąskiej AI (wykonującej jeden typ zadań), sztucznej ogólnej inteligencji (wiele typów zadań) oraz ogólnej sztucznej inteligencji (AGI, ang. *Artificial General Intelligence*; każdy typ zadań). Jednak w kolejnych częściach nie odwołują się do tych rozróżnień, pozostając przy ogólnym określeniu AI. Podkreślają, że trudno stwierdzić czy duże modele językowe (LLM, ang. *Large Language Models*) faktycznie rozumieją wycinki rzeczywistości, co do których udzielają trafnych odpowiedzi. Konkluzja jest dość oczywista i brzmi następująco: pomimo ogromnego postępu AI nadal posiada liczne braki względem ludzkiej inteligencji, m.in.: brak zdrowego rozsądku (tj. rozumienia i rozumowania o podstawowych czynnościach); zazwyczaj radzi sobie tylko z jednym rodzajem zadań; brak umiejętności tworzenia wieloetapowych planów działań i przechodzenia od planów ogólnych do szczegółowych; brak rozumienia emocji, zależności społecznych i kulturowych; brak umiejętności uczenia się na podstawie nielicznych przykładów; niska umiejętność interakcji z fizycznym światem; wysoka podatność na zmiany kontekstu; trudność w interpretacji wykonywanych działań (s. 18–33).

Rozdział 2. pt. „Czy AI może być bezpieczna?" zaczyna się od tezy, że AI niesie ryzyka, podobnie jak inne wytwory techniki. Największą obawą jest teoretyczna możliwość zagłady ludzkości. AI zwiększanie tylko nasze możliwości, lecz także pomaga ludziom chcącym szkodzić innym przez ataki phishingowe, farmy botów ideepfake. Niedostatki AI już teraz powodują szkody. AI w wojsku wzbudza kontrowersje, bo zwiększa skuteczność ataków i ogranicza liczbę przypadkowych ofiar, ale też może ułatwiać działania terrorystom. Media społecznościowe stosują wiele rodzajów AI, aby zatrzymać naszą uwagę jak najdłużej i serwować nam reklamy. Odpowiedź na tytułowe pytanie rozdziału mogłaby brzmieć następująco (gdyby autorzy wyrazili ją jednoznacznie): AI,

jak każde narzędzie, może przynieść wiele pożytku, ale też wyrządzić szkody. Na chwilę obecną wszystko zależy od jej użytkowników, gdyż sama AI nie potrafi oceniać wykonywanych działań pod kątem moralnym.

Rozdział 3. pt. „Czy AI może szanować prywatność?" przytacza kazus DeepNude, tj. AI stworzonej do „rozbierania" osób na zdjęciach. To przykład uprzedmiotowienia, naruszenia prywatności i godności. Autorzy przytaczają kilka definicji „prywatności", podkreślając, że jest to wartość, którą należy respektować, a każdy człowiek ma do niej prawo. AI może naruszać to prawo poprzez odbieranie kontroli nad informacją nas dotyczącą. Dzieje się totakże przez rozpoznawanie twarzy w monitoringu miejsc publicznych, zwłaszcza podczas zebrań, protestów czy zgromadzeń religijnych – nie zawsze chcemy, by odnotowano naszą obecność. Autorzy wskazują, że wymaganie ogromnych ilości danych do rozwijania AI stanowi zagrożenie pośrednie, ponieważ motywuje do gromadzenia danych w celu ich sprzedaży firmom zajmujących się rozwojem tej technologii. Jednym ze sposobów ochrony prywatności (tj. kontroli informacji) może być regulamin dotyczący przetwarzania danych. W praktyce takie rozwiązanie jest nieefektywne (tekst jest długi i niezrozumiały, więc niemal nikt go nie czyta) oraz nieskuteczne (niewyrażenie zgody może oznaczać niemożliwość korzystania z serwisu bądź aplikacji, które nie mają alternatywy). Innym popularnym sposobem jest tzw. anonimizacja, mająca uniemożliwić jednoznaczną identyfikację osób, których dane znajdują się w bazie. Jednak korelacja danych z różnych baz pozwala na identyfikację niemal wszystkich (s. 102–104). Rozdział kończy się tezą, że podejmowane są działania mające zmniejszyć skalę naruszeń oraz że rozwój AI nie stoi w sprzeczności z zachowaniem prywatności.

Rozdział 4. pt. „Czy AI może być sprawiedliwa?" dotyczy uprzedzeń wobec pewnych grup, np. osób ciemnoskórych, ubogich, imigrantów i kobiet. Uprzedzenia są wynikiem stereotypów i nierównej reprezentacji tych grup w zbiorach danych, użytych w tworzeniu AI. Autorzy wprowadzają trzy pojęcia sprawiedliwości: dystrybutywną (podział obowiązków i korzyści), retrybutywną (adekwatność kar do win) i proceduralną (uczciwość procesu). AI jest używane w sądownictwie amerykańskim do oceny czy oskarżony powinien czekać na proces w areszcie, czy też zostać zwolniony. Ocena opiera się na określeniu prawdopodobieństwa niestawienia się na rozprawie i/lub popełnienia kolejnego przestępstwa. Problemem jest brak informacji o przyjętej definicji „sprawiedliwości" w takich systemach. AI, podobnie jak sędziowie, posiada niedoskonałości i uprzedzenia, co narusza sprawiedliwość dystrybutywną. Autorzy wskazują, że ważne jest rozstrzygnięcie czy sugestie AI naruszają sprawiedliwość retrybutywną, ale nie rozwijają tego wątku. Wykorzystanie AI narusza także sprawiedliwość proceduralną, ponieważ proces rekomendacji jest skomplikowany i niezrozumiały. W praktyce nie ma możliwości podważenia

zaleceń ani odwołania się od nich, więc proces jest nieuczciwy. Rozdział kończy się tezą, że trwają prace nad eliminacją uprzedzeń oraz „interpretowalną AI", tj. takim projektowaniem modeli AI, aby ich sposób podejmowania decyzji był możliwy do zrozumienia przez ludzi.

Rozdział 5. pt. „Czy AI (lub jej twórcy, lub użytkownicy) może być odpowiedzialna?" dotyczy odpowiedzialności rozumianej jako podleganie sankcjom moralnym (ostracyzm społeczny, poczucie wstydu i winy) i prawnym (grzywny i pozbawienia wolności). Pytanie o odpowiedzialność za szkody wyrządzone przez AI jest pytaniem o to kto podlega sankcjom. Autorzy przytaczają kazus z 2018 roku, gdy testowany w Arizonie, przez firmę Uber, autonomiczny samochód śmiertelnie potrącił pieszą, przechodzącą przez ulicę w niedozwolonym miejscu. Analizują argumenty za możliwą odpowiedzialnością lub jej brakiem: A. kobiety w fotelu kierowcy (nie patrzyła na drogę przez kilka sekund przed wypadkiem, choć ten wydarzył się na tyle szybko, że mogłaby nie zdążyć zareagować); B. pieszej (była ciemno ubrana, pod wpływem narkotyków i przechodziła w niedozwolonym miejscu); C. twórców AI (nie wzięli pod uwagę, że ludzie mogą przechodzić w miejscach niedozwolonych i popełnili kilka innych błędów w projekcie); D. firmy Uber (niskie standardy bezpieczeństwa; ignorowanie wcześniej występujących problemów); E. rządu Arizony (brak nadzoru nad standardami bezpieczeństwa; dopuszczenie do testów pojazdów firmy Uber, choć wcześniej rząd Kalifornii zabronił ich u siebie); F. AI (na obecnym etapie rozwoju AI nie podlega sankcjom prawnym; możliwe, że do pewnego stopnia dotyczą jej sankcje moralne). Autorzy nie podejmują się rozstrzygnięć, lecz wskazują, że brak rozwiązań rozmywa kwestię odpowiedzialności, co skutkuje brakiem zachęt do poprawy skuteczności systemów, problemami z kompensacją szkód oraz zrzucaniem odpowiedzialności na samą AI.

W rozdziale 6. pt. „Czy AI może przyjąć ludzką moralność?", autorzy proponują „wbudowanie" ludzkiej etyki w AI, zamiast tworzenia „etyki AI", aby jej działania były bezpieczniejsze i dopasowane do „naszych" wartości. Na podstawie własnych badań proponują stworzenie AI wspomagającej decyzje dotyczące przeszczepów. Zaczęli od określenia istotnych cech (*features*) poprzez ankiety z otwartymi pytaniami zadawanymi możliwie dużej i zróżnicowanej grupie respondentów. Na tej podstawie wskazano listę cech, które inna grupa oceniła pod kątem istotności (wagi, hierarchii) moralnej poprzez wybór osoby mającej otrzymać przeszczep (s. 175–176). Potem stworzono statystyczny model dokonanych ocen. Autorzy twierdzą, że stosowali wyłącznie metody wyjaśnialnej AI, pozwalające określić wpływ poszczególnych cech na podjęte decyzje. AI z ludzką moralnością będzie popełniać „ludzkie" błędy i z tego względu powinna odzwierciedlać pewien ideał, a nie rzeczywisty proces. Autorzy proponują przeprowadzenie testu, ocenę wiedzy respondenta na temat przeszczepów, uzupełnienia braków wiedzy i ponowne przeprowadzenie testów

(s. 181–183). Tak chcą oszacować wpływ niewiedzy na decyzje i dostosować algorytm. Podkreślają, że system bazujący na idealizacji nie będzie bezbłędny, lecz nadal będzie lepszy niż systemy pozbawione moralności bądź oparte o rzeczywiste ludzkie osądy. Autorzy określają go jako „sztuczna ulepszona demokracja" (*artificial improved democracy*), a jego celem ma być pomoc ludziom oraz AI w dokonywaniu lepszych moralnych sądów (s. 186). Wydaje się jednak, że ten projekt idealizacji zbyt upraszcza problem. Wiele badań dowodzi, że wpływ na wyniki ankiet ma sformułowanie i kolejność zadawania pytań. Sposób przekazywania wiedzy, która ma być uzupełniana, również ma wpływ na uzyskane później odpowiedzi. Tak utworzony model moralności AI będzie niemożliwy do odtworzenia (problem replikacji) oraz będzie ściśle zależny od uwarunkowań kulturowych na konkretnym obszarze. Głoszona przez autorów wyższość wyidealizowanego osądu moralnego (s. 185) niejawnie zakłada, że istnieje jedna prawdziwa etyka. To założenie stoi za tezą, że takie oceny będą lepsze niż oceny przyjmowane w oparciu o nieulepszony model. Jednak decyzja, że nerkę powinien otrzymać pacjent „A" nie jest lepsza ani gorsza od decyzji, że powinien ją otrzymać pacjent „B". Moim zdaniem wybór między „A" i „B" jest fałszywym dylematem. Można powiedzieć, że każdy z tych wyborów jest nieskończenie lepszy niż niepodejmowanie żadnego działania. Odwoływanie się do zasad i wartości dostarcza przede wszystkim racjonalizacji podejmowanych działań, a ocena jest dokonywana w zależności od przyjętych kryteriów odzwierciedlających społeczne oczekiwania.

Rozdział 7. pt. „Co możemy zrobić?" stawia pytanie: „co musimy zrobić, aby upewnić się, że wpływ AI na nasze życie jest etyczny?" [sic!] (s. 189). Autorzy zauważają rozłam między zasadami dotyczącymi moralnej AI a faktycznie powstającą AI. Przyczyną jest brak obowiązku przestrzegania zasad (nie są one prawem) ale także brak treningu twórców AI, aby podejmować decyzje mające moralne i społeczne implikacje. Narzędzia wspomagające tworzenie moralnej AI nie uwzględniają braku komunikacji między jej twórcami i użytkownikami oraz presji czasowej i finansowej w firmach tworzących AI. Twórcom zależy na stworzeniu produktu, który szybko będzie dochodowy, więc kwestie związane z etyką AI są na drugim planie. A po utworzeniu produktu zajmują się jego rozwijaniem, zamiast modyfikacjami w celu rozwiązania problemów etycznych. Ponadto, nie ma ogólnie przyjętej metryki pozwalającej monitorować „moralny wpływ" procesów i produktów opartych o AI. Autorzy przedstawiają pięć wezwań do działania: 1. zwiększaj skalę narzędzi do rozwoju moralnej AI; 2. dziel się wiedzą dotyczącą dobrych praktyk wdrażania moralnej AI; 3. zapewnij szkolenia związane z etyką AI; 4. podejmij dyskusję z użytkownikami; 5. wdrażaj publiczne zasady dotyczące moralnej AI. Rozdział kończy się twierdzeniem, że nie da się stworzyć AI bez przyjmowania założeń na temat tego co jest dobre i złe moralnie, ponieważ działania każdego systemu mają

moralnie istotne skutki. Zatem moralna AI obejmuje świadome i przemyślane podejmowanie decyzji, które będą miały moralnie istotne konsekwencje. Nie jest to kwestia wyłącznie techniczna ani wyłącznie etyczna.

We wnioskach autorzy podkreślają, że rozwój AI, jak każdej innej technologii, przynosi korzyści oraz problemy. Nie sposób przewidzieć jakie spowoduje szkody. Nie oznacza to, że nie powinniśmy próbować oceniać jej wpływu i ryzyka. Ludzka zdolność moralnej oceny jest zawodna, szczególnie w przypadku technologii, która rozwija się szybko i przynosi jej twórcom ogromne zyski. Nie możemy pozwolić, by postęp zaabsorbował naszą uwagę tak, że przeoczymy etyczne problemy bądź uznamy, że nic nie można zrobić. Autorzy kończą optymistycznym stwierdzeniem, że poprzez prowadzenie badań, tworzenie regulacji i podejmowanie działań edukacyjnych nadal mamy wpływ na to jak AI jest kształtowana.

Książka napisana jest jasnym językiem, lecz poszczególne części nie są równe: najkrótsza liczy 25 stron, a najdłuższa 41. Postawiony w przedsłowiu cel to przedstawienie problemów oraz możliwych do realizacji sugestii (*action ablesuggestions*). Można go uznać za zrealizowany w rozdziale 7., gdzie autorzy zauważają rozłam między zasadami dotyczącymi moralnej AI i praktyką tworzenia technologii oraz przedstawiają pięć wezwań do działania (*call to action*). Co prawda „wezwanie" jest czymś innym niż „sugestia", jednak pozostałe części książki nie zawierają konkretnych wskazówek odnoszących się do prowadzenia rozwoju AI.

Większość poruszanych zagadnień jest omówiona pobieżnie i bez wyraźnego związku z głównym tematem. Brakuje *explicite* wyrażonego połączenia między nimi a moralną AI, np. wskazania, że spośród rodzajów AI przedstawionych w 1. rozdziale tylko AGI może osiągnąć moralną podmiotowość, a reszta jest wyłącznie narzędziem. Ze względu na powierzchowność tekst jest niejednoznaczny: Autorzy, pisząc o „moralnej AI", nie rozróżniają czy piszą o narzędziu, jego wykorzystaniu, czy procesie tworzenia. Sam fakt występowania uprzedzeń w zbiorach danych jest moralnie neutralny, choć jest błędem poznawczym. Trenowanie AI w oparciu o błędne założenia lub źle dobrane dane podlega ocenie moralnej. Jeśli jest to wynik świadomego działania bądź zaniedbania, to wiąże się z moralną winą i wymaga korekty. Natomiast, jeśli nie było to zamierzone, to także wymaga korekty, choć może nie wiązać się z moralną winą. Jednak wydaje się, że autorzy skupiają się bardziej na ewaluacji działań wykonywanych przy użyciu AI.

Najważniejszą częścią pod względem treści jest najkrótszy (25 stron) rozdział szósty. Zawiera nowatorski pomysł utworzenia „demokratycznej etyki" oraz opis propozycji badań, które w pewnym stopniu były przez autorów prowadzone. Niestety, problematyce „wcielenia" moralności w AI poświęcono ledwie 15 stron (s. 173–187). Rozdziały 1–5 wydają się być wprowadzeniem

przez luźno powiązane dygresje dające pretekst do krótkiego przedstawienia własnych badań (r. 6) oraz wezwań kierowanych do twórców AI (r. 7).

Największą słabością tekstu jest powierzchowność uniemożliwiająca adekwatne omówienie poruszonych zagadnień. Autorzy nie podjęli pytań dotyczących natury AI, których rozwinięcie zapowiedzieli we wstępie (s. XVIII–XIX). Ponadto, nie zauważają, że porównują obecne i spodziewane możliwości AI do możliwości ludzkich na bardzo ogólnym poziomie oraz na podstawie uproszczonego obrazu człowieka i samej AI. Przykładowo, porównywana w pierwszym rozdziale (s. 24–25) zdolność empatii, rozumienia relacji społecznych i kulturowych z jednej strony ściśle zależy od danej kultury (te same gesty niosą różne znaczenia), a z drugiej nie występuje tak samo u wszystkich – osoby ze spektrum autyzmu często mają problem z adekwatnym odczuwaniem empatii oraz odnalezieniem się w społeczeństwie i kulturze, choć bez wątpienia są ludźmi. Rodzi to pytanie, których ludzi AI ma odwzorowywać i z kim należy ją porównywać? Czy wystarczy, że będzie posiadać cechy niektórych ludzi, czy też ma odzwierciedlać pewien ogólny ideał człowieka?

Autorzy, mimo wprowadzonych rozróżnień, w całym tekście stosują potoczną narrację. Nie odróżniają działań AI jako narzędzia od użycia narzędzia przez człowieka. Wielokrotnie piszą o AI jakby była świadomym i autonomicznym podmiotem, choć zaznaczyli (s. 2–3), że nie podejmują się rozstrzygnięcia tych kwestii. Ponadto, błędnie stosują pojęcie „etyczny", np. na stronie 189, gdzie pytają „co trzeba zrobić, aby wpływ AI na nasze życia był etyczny". Z kontekstu wynika, że powinno się tu znaleźć słowo „korzystny". Prezentowane w tekście podejście bardzo upraszcza problematykę etyczną – autorzy nie biorą pod uwagę, że rozwój AI, jak każdej innej technologii, może spowodować znaczące zmiany w naszym codziennym życiu, a jednocześnie być moralnie neutralny. Proponowane przez nich podejście utworzenia etyki „demokratycznej" jest nietypowe i wymaga więcej objaśnień oraz opracowania – dlaczego takie podejście? w czym jest lepsze od innych?

Niestety, autorzy nie wskazują do kogo kierowana jest recenzowana książka. Układ treści i zawarte w niej podstawowe pojęcia, wraz z bardzo ogólnym opisem działania niektórych rodzajów AI, mogłyby wskazywać, że przeznaczona jest dla mniej zaawansowanych czytelników, jednak zawarte w rozdziale 7. wezwania są wprost kierowane do twórców AI, którzy z pewnością nie potrzebują omówienia podstaw. Ze względu na powierzchowność rozdziałów 1–5 bardziej zaawansowani czytelnicy nie znajdą tam zbyt wielu nowych informacji. Rozdział 6. zawiera propozycję utworzenia „demokratycznej etyki" oraz opis proponowanych i już wykonanych badań, co w połączeniu z zawartym we wnioskach stwierdzeniem, że poprzez badania, regulacje i edukację nadal możemy mieć wpływ na rozwój AI, pozwala sądzić, że te fragmenty są kierowane do naukowców. Jednak trudno stwierdzić do kogo kierowana jest całość omawianego tekstu.

Zatem, książka fragmentami będzie dobrą propozycją dla mniej zaawansowanych czytelników, którzy będą mieli okazję do poszerzenia wiedzy. Natomiast pozostali mogą skupić się na dwóch ostatnich rozdziałach, zawierających opis propozycji stworzenia moralnej AI oraz wezwania do praktycznych działań.