

Toruń, 27.01.2026.

Recenzja rozprawy doktorskiej mgr. Macieja MROZA nt. „Morality of Artificial Agents” („Moralność sztucznych agentów”), s. 171, wykonanej pod kierunkiem prof. dr. hab. Krzysztofa Stachewicza w dziedzinie nauk teologicznych, w dyscyplinie nauki teologiczne na Wydziale Teologicznym Uniwersytetu im. Adama Mickiewicza w Poznaniu w 2025 r.

Niniejszą recenzję przedkładam w odpowiedzi na uchwałę nr 207/RD_NT/WT/2025 Rady Naukowej Dyscypliny nauki teologiczne Uniwersytetu im. Adama Mickiewicza w Poznaniu z dnia 20 października 2025, na mocy której zostałem wyznaczony na recenzenta rozprawy doktorskiej Pana mgr. Macieja MROZA, pt. „Morality of Artificial Agents”, przygotowanej pod kierunkiem prof. dr. hab. Krzysztofa Stachewicza.

Uwagi wstępne

Według raportu IDEAS NCBR „Zatrzymać najlepszych – trendy w kształceniu w obszarze AI na poziomie doktoranckim”, pod redakcją dr Grażyny Żebrowskiej, przygotowanego we współpracy z Ośrodkiem Przetwarzania Informacji – Państwowym Instytutem Badawczym, od roku 2020 do połowy 2023 obroniono w Polsce 207 doktoratów związanych z problematyką sztucznej inteligencji w dyscyplinach informatycznych i pokrewnych. Oznacza to średnio ok. 50 doktoratów rocznie w obszarze AI spośród 4,7 tys. ogółem obronionych doktoratów. Większość prac dotyczyła technicznych aspektów sztucznej inteligencji. Z publicznie dostępnych raportów nie wynika, ile prac doktorskich poruszało etyczne aspekty sztucznej inteligencji i autonomicznych agentów. Niestety nie ma potwierdzonych statystyk, ile z tych prac poruszało zagadnienia z obszaru etyki sztucznej inteligencji (*AI Ethics*). Stwierdzony stan rzeczy pozwala wysunąć przypuszczenie, że rozprawa doktorska mgr. Macieja Mroza najprawdopodobniej jest jedną z pierwszych tego rodzaju prac w Polsce. Oznacza to, że

wypełnia ona istotną lukę w rodzimych badaniach naukowych nad sztuczną inteligencją, prowadzonych z pozycji etycznych.

Ponieważ recenzowana praca została wykonana w dziedzinie nauk teologicznych, w dyscyplinie nauki teologiczne, usiłowałem sprawdzić, czy oraz ewentualnie ile i jakie rozprawy w tym obszarze powstały w Polsce. Niestety z uwagi na to, że repozytoria uczelniane w Polsce są rozproszone – nie ma jednej centralnej bazy doktoratów z precyzyjnymi tematami, a także w związku z tym, że doktoraty tego rodzaju mają charakter interdyscyplinarny, co powoduje, że prace łączące teologię i AI często formalnie klasyfikowane są w ramach szerokich dyscyplin (takich, jak etyka, filozofia czy nauki społeczne), a nie wyraźnie jako „teologia sztucznej inteligencji”, nie byłem w stanie ustalić, czy istnieją podobne do dysertacji mgr. Mroza rozprawy. Mimo to, sądząc po ogólnej liczbie prac doktorskich w Polsce, poruszających problematykę sztucznej inteligencji, oraz biorąc pod wzgląd wciąż nieliczne publikacje naukowe w Polsce, których autorzy podejmują tę problematykę, osadzając ją w kontekście etycznym, można stwierdzić, że dysertacja mgr. Mroza należy do pionierskich osiągnięć naukowych.

O ile w teoretycznych opracowaniach międzynarodowych etyka sztucznej inteligencji ma dobrze rozpoznane podziały sekcyjne: etyka techniczna, nazywana etyką maszynową lub komputacyjną, etyka stosowana, etyka odpowiedzialności, etyka regulacyjna, etyka społeczna, metaetyka AI, to w Polsce obserwowalne tendencje badawcze wskazują na pewne dominujące obszary oraz inne, jeszcze słabiej reprezentowane. Na podstawie przeglądu polskich publikacji i przedsięwzięć naukowych można wyróżnić następujące dominujące typy etyki AI w Polsce oraz te, które są rzadziej poruszane. Są to: etyka praktycznych zastosowań AI, etyka regulacyjna i prawna oraz etyka społeczno-kulturowa. Stosunkowo rzadko reprezentowane są w Polsce następujące obszary badań w zakresie etyki sztucznej inteligencji: etyka sztucznych agentów moralnych (*machine ethics*), metaetyka AI i ontologiczne aspekty działania AI, etyka projektowania AI (*ethics by design, value-sensitive design*).

Z satysfakcją odnotowuję, że praca doktorska mgr. Macieja Mroza w sposób kompleksowy porusza bez mała wszystkie wyszczególnione aspekty etyki sztucznej inteligencji. Czyni to w sposób kompetentny, przejrzysty, spójny i konkluzywny.

Analiza celów, struktury oraz metod zastosowanych w pracy

Cel badawczy

Przedstawiony we wprowadzeniu (s. 6) cel pracy został enigmatycznie określony jako „wniesienie wkładu do dyskursu naukowego w obszarze etyki AI, przy zachowaniu perspektywy teologicznej”. Należałoby w tym miejscu wyjaśnić, jaki konkretnie wkład

recenzowana praca wnosi do dziedziny etyki sztucznej inteligencji, jak również, jak pojmuję się w niej tę dziedzinę. Owszem, w dalszych partiach pracy jest mowa o tym, jakie etyka sztucznej inteligencji przyjmuje urzeczywistnienia w literaturze naukowej. Trudno jednak stwierdzić, czy doktorant opowiada się za zbiorczym rozumieniem tej etyki, czy też wyraźnie orientuje się na określony jej aspekt. Na podstawie tego doktoratu da się dość precyzyjnie określić, jaki rodzaj etyki AI jest w nim dyskutowany, mimo że nie został on nazwany w nim wprost jednym terminem. Co więcej, można również wskazać, jakich rodzajów etyki AI doktorant świadomie unika lub marginalizuje. Syntetycznie rzecz biorąc, z doktoratu nie wynika, aby autor zagłębiał się – co nie znaczy, że zupełnie pomija – w takie rejony etyki sztucznej inteligencji, jak etyka regulacyjna (*governance AI*) – przedstawiane dokumenty omawiane są deskryptywnie, a nie normatywnie ani projektowo; etyka stosowana, oparta na metodzie *case-based reasoning* (brak jest w pracy analizy konkretnych zastosowań *AI Ethics* w tak newralgicznych obszarach, jak medycyna, *policing* czy *Human Resources*; brak jest pogłębionych studiów przypadków, choć konkretne wskazania na takowe, rzecz jasna, występują; czy rozszerzającej dyskusji konfliktów norm i rozwiązań praktycznych); etyka odpowiedzialności systemowej (*accountability ethics*) – zagadnienia odpowiedzialności prawnej, organizacyjnej, korporacyjnej wydają się w pracy świadomie pomięte. Z uwagi jednak na wyraźnie filozoficzny charakter pracy jest to akurat krok w dobrą stronę.

Poczyniona uwaga jest o tyle istotna, że to właśnie te przestrzenie tematyczne etyki sztucznej inteligencji dominują dziś w Polsce, tym czasem – dodałbym: na szczęście – rozprawa mgr. Mroza podąża inną, własną drogą. Jaka to droga? Jest trzon stanowi perspektywa metaetyczna i ontologiczna etyki sztucznej inteligencji, skoncentrowana na problemie moralnej sprawczości (*moral agency*). Mamy tu więc do czynienia z doktoratem z obszaru *Ethics of Artificial Moral Agency* (etyki sztucznej moralnej sprawczości). Centralnym pytaniem pracy o fundamentalnym znaczeniu dla tej dziedziny wiedzy jest kwestia: czy i w jakim sensie sztuczne systemy mogą być w sposób uzasadniony kwalifikowane jako moralne podmioty. Tak określony przedmiot dociekań lokuje rozprawę jednoznacznie w nurcie *machine ethics* (w sensie filozoficznym, nie inżynierskim). Nie chodzi w niej jednak o wskazanie gotowych rozwiązań w zakresie preskrypcji norm etycznych, określających normatywne ramy dla funkcjonowania sztucznych agentów, ich projektowania i zastosowań. Praca doktorska mgr. Mroza jest do imentu filozoficzna, a przy tym solidnie osadzona w międzynarodowej, interdyscyplinarnej, głównie anglojęzycznej (co oczywiste, zważywszy na ledwie raczkujące w Polsce badania i publikacje dotyczące sztucznej inteligencji), literaturze naukowej. Doktorant unika z jednej strony pułapki nadmiernego uteoretyzowania, ocierającego się o sofistyczne przeintelektualizowanie podejmowanego tematu,

a z drugiej nadmiernego przeciążenia pracy technikami. Szczęśliwie doktorant wystrzega się kroczenia drogą, znaną choćby z bioetyki czy psychologii moralności, na których spoczywają niezmiennie dwa podstawowe zarzuty: nadmiernego przeintelektualizowania w jednym przypadku oraz nadmiernej medykalizacji dyskursu w drugim. Doktorat mgr. Macieja Mroza dotyczy zasadniczej filozoficznej kwestii, której doniosłość przekłada się, jako żywo, na rozstrzygnięcia natury normatywnej, proceduralnej, projektowej, praktycznej, społecznej, politycznej i gospodarczej. Otóż Maciej Mróz dąży w swojej pracy przede wszystkim do ustalenia tego, jakie są warunki możliwości uzasadnionego naukowo przypisania sztucznej inteligencji podmiotowości i sprawczości moralnej.

Skoro jednak o etyce mowa, to zachodzi zrazu pytanie: w jakim znaczeniu doktorant operuje w swojej pracy terminem ‘etyka’? Czy traktuje go jako równoważny użytemu w tytule rozprawy terminowi ‘moralność’? Jakie jest uzasadnienie dla literalnie zrozumiałego, acz arbitralnie sformułowanego stwierdzenia, że [k]ontekst sztucznej inteligencji wymaga podejścia wielodyscyplinowego? Na czym doktorant opiera przekonanie, że właśnie takie podejście pozwoli okiełznać „terminologiczną niestabilność” i „rozbieżne sposoby użycia kluczowych” dla dyskursu naukowego dotyczącego AI terminów? Otóż wydaje mi się, że w tej materii jest zgoła odwrotnie. Multidyscyplinaryzacja tego dyskursu prędzej wzmacnia tendencję do zaciemniania i destabilizacji dyskursu naukowego, niż ją ogranicza. Interesująca byłaby również odpowiedź na pytanie, dlaczego doktorant problematykę etyki AI usiłuje koniecznie „osadzić w perspektywie teologicznej”? Czy to jedynie zabieg formalny, czy też doktorant jest może zdania, że dopiero teologia pozwala skutecznie i sensownie wyjaśnić, uzasadnić lub rozstrzygnąć rozbieżności i aporie, napotykane na gruncie etyki sztucznej inteligencji? Podany na 8. stronie pracy przez doktoranta argument, że „sztuczna inteligencja staje się przedmiotem coraz większego zainteresowania teologii”, jest trafny i na swój sposób przekonujący. Czy jednak jest on wystarczający do uzasadnienia zaprzężenia perspektywy teologicznej do tematu na wskroś filozoficznego? Znacznie bardziej przekonującą, w tym aspekcie, racją, jest stwierdzenie, że choć AI nie jest obecnie identyfikowane wprost jako *locus theologicus*, to jednak jako zjawisko kulturowo-techniczne sztuczna inteligencja może być i w coraz większym stopniu staje się ważnym przedmiotem refleksji teologicznej w ramach *loci theologici* związanych z kulturą i antropologią.

Struktura pracy

Rozprawa doktorska Macieja Mroza posiada klarowną, logicznie uporządkowaną i uzasadnioną metodologicznie strukturę, podporządkowaną głównemu problemowi

badawczemu, którym jest możliwość przypisania sztucznej inteligencji statusu podmiotu moralnego oraz uzasadnienie sensowności atrybucji pojęcia moralności do sztucznych agentów, realizujących „inteligentne” zadania.

Praca składa się ze wstępu, czterech rozdziałów merytorycznych oraz zakończenia. Całości dopełnia obszerna, zorientowana interdyscyplinarnie i nader aktualne, gdyż sięgające pozycji nawet z roku 2025, zestawienie bibliograficzne. Układ rozprawy ma charakter problemowo-systematyczny, a nie historyczno-autorski, co odpowiada deklarowanej w niej metodologii badawczej.

We *Wstępie* autor precyzyjnie nakreśla kontekst badawczy, sytuując problematykę sztucznej inteligencji na styku filozofii, etyki i teologii, oraz formułuje cele pracy. Zarysowana zostaje perspektywa metodologiczna, obejmująca analizę krytyczną współczesnych debat nad AI, *machine ethics* oraz *Artificial Moral Agents*, przy jednoczesnym zachowaniu antropologicznego i teologicznego horyzontu refleksji. Wstęp zawiera: uzasadnienie wyboru tematu, określenie zakresu pracy, wyjaśnienie przyjętych metod badawczych oraz zastrzeżenia dotyczące granic prowadzonych analiz (np. marginalizacja zagadnień prawnych i regulacyjnych).

Rozdział pierwszy (zatytułowany „*Navigating the Complex Mosaic of Artificial Intelligence*” – „Poruszanie się po złożonej mozaice sztucznej inteligencji”) ma charakter wprowadzająco-porządkujący. Doktorant rekonstruuje w nim podstawowe sposoby definiowania sztucznej inteligencji, ukazując ich historyczne, techniczne i filozoficzne uwarunkowania. Omawiane są tu zarówno klasyczne ujęcia Alana Turinga, Stuarta Russella i Petera Norviga, jak i współczesne perspektywy (AGI – *Artificial General Intelligence* [Ogólna Sztuczna Inteligencja], LLM – *Large Language Model* [Duży Model Językowy], krytyka „stochastycznych papug” oraz ujęcia regulacyjne i społeczno-techniczne. Celem rozdziału jest precyzacja i ustabilizowanie aparatu pojęciowego, ukazanie wieloznaczności terminu „AI” oraz przygotowanie gruntu pod dalszą analizę problemu sprawczości i moralności. Rozdział ten nie zawiera jeszcze własnych rozstrzygnięć normatywnych doktoranta. Pełni on jednak ważną rolę konceptualnego zaplecza dla dalszych wywodów i rozwinięć.

Rozdział drugi dotyczy filozoficznych podstaw i współczesnych debat toczonych wokół zagadnienia sztucznej sprawczości (*Artificial Agency*). Stanowi on kluczowe ogniwo teoretyczne rozprawy. Doktorant przechodzi w nim od pojęcia inteligencji do pojęcia sprawczości, rekonstruując klasyczne i współczesne filozoficzne koncepcje sprawczości, zarówno w odniesieniu do podmiotów ludzkich, jak i systemów sztucznych. W rozdziale tym omówione zostają m.in. definicyjne kryteria pojęcia sprawczości; zagadnienie stopniowalności i progów sprawczości; relacje między sprawczością, osobowością i odpowiedzialnością, a także pojęcia *patency* (bycia podmiotem moralnego

traktowania, czyli odbiorcą działań moralnych) i podatności na krzywdę. Rozdział ten ma zasadniczy charakter dla całej rozprawy, ponieważ wprowadzone w nim zostają kryteria, które w dalszej części pracy znajdują zastosowanie do oceny możliwości istnienia sztucznych podmiotów moralnych.

Rozdział trzeci, pt. „*Machine Ethics and Artificial Moral Agents* („Etyka maszyn i sztuczne podmioty moralne”) ma charakter systematycznej rekonstrukcji debaty, toczonej w kręgach naukowych, nad etyką maszyn. Doktorant omawia w nim główne stanowiska dotyczące projektowania sztucznych agentów moralnych, w tym klasyfikacje *Artificial Moral Agents*, podejścia *top-down* (odgórne), *bottom-up* (oddolne) i hybrydowe, problem testów moralnych (np. *moral Turing test*), spór między zwolennikami i krytykami budowy *Artificial Moral Agents* (sztucznych agentów moralnych), paradygmat *value alignment* (współmierności implementowanych do systemów AI wartości z wartościami ludzkimi) oraz problem kontroli sztucznej inteligencji. Rozdział ten ma charakter wyraźnie deskryptywno-analityczny, ale zawiera także elementy krytyczne, które zapowiadają rozstrzygnięcia przedstawione w kolejnym, ostatnim już rozdziale rozprawy. Przyjmuje on postać mapy stanowisk, umożliwiającej osadzenie własnych tez i rozstrzygnięć doktoranta pośród istniejącej literatury naukowej.

Niezwykle interesujący rozdział czwarty mierzy się z arcyważnym pytaniem: „*Can Morality Be Computed?*” („Czy moralność daje się zalgorytmizować?”). Rozdział ten ma charakter syntetyczny i krytyczny. Stanowi on zasadnicze miejsce formułowania przez doktoranta zrębów jego własnego stanowiska. Przedmiotem analizy są tutaj dylematy związane z możliwością redukcji moralności do procesów obliczeniowych oraz adekwatność modeli funkcjonalistycznych w opisie aktu moralnego. W rozdziale tym doktorant konfrontuje podejścia redukcjonistyczne z ujęciami nieredukcyjnymi; rozpatruje LLM – *Large Language Model* [Duże Modele Językowe] w kategoriach „skompresowanych modeli świata”; wprowadza pojęcie kreatywnego charakteru aktu moralnego oraz wskazuje na fundamentalne ograniczenia projektów etyki maszyn. Rozdział ten stanowi nader kompetentne i z powodzeniem zrealizowane argumentacyjne zwieńczenie pracy.

W *Zakończeniu* doktorant syntetyzuje wyniki przeprowadzonych analiz, wskazując na ich konsekwencje dla możliwych kierunków i efektów dalszej debaty naukowej nad moralną sprawczością AI, etyką sztucznej inteligencji oraz szerszym kontekstem technospołecznym i antropologicznym. Zakończenie, co oczywiste, nie wprowadza nowych wątków, lecz porządkuje ustalenia i wskazuje możliwe kierunki dalszych badań.

Metodologia pracy

W pracy zastosowano metody *stricte* filozoficzne, uzupełnione analizą technicznego i etycznego dyskursu AI. Doktorant nie prowadzi ani nie sprawozdaje własnych badań empirycznych ani eksperymentów, tylko pracę analityczno-krytyczną. Zestaw metod wykorzystywanych w rozprawie został jasno zarysowany we Wstępie. Najważniejsze z nich to: metoda egzegetyczna (rekonstrukcyjna), metoda krytyczna (dominująca zwłaszcza w rozdziałach 3 i 4), konsekwentnie stosowana w całej pracy metoda analizy pojęciowej, metoda porównawcza, elementy metody fenomenologicznej, metoda hermeneutyczna i analiza źródeł oraz metoda syntetyczna.

Język dysertacji oraz walory warsztatowe

Szata językowa rozprawy doktorskiej Macieja Mroza odznacza się wysokim poziomem komunikatywności, poprawności i precyzji sformułowań oraz godnej podkreślenia dojrzałości akademickiej. Doktorant sprawnie posługuje się klarownym, dobrze zorganizowanym stylem naukowym, adekwatnym do charakteru pracy filozoficzno-teologicznej, podejmującej zagadnienia z pogranicza etyki, filozofii techniki oraz teorii sztucznej inteligencji. Wywód prowadzony jest z żelazną konsekwencją, w sposób logiczny oraz z zachowaniem należytej dyscypliny terminologicznej.

Na szczególne podkreślenie zasługuje fakt, że dysertacja została napisana w języku angielskim na poziomie w pełni odpowiadającym standardom międzynarodowego dyskursu akademickiego. Doktorant swobodnie i rzetelnie posługuje się specjalistycznym słownictwem zarówno z zakresu filozoficznej etyki, jak i współczesnych badań nad sztuczną inteligencją. Terminologię naukową stosuje w sposób transparentny, uzasadniony i konsekwentny. Ewentualne wieloznaczności pojęć (takich, jak np. „wartość”, „sprawczość”, „inteligencja”, „autonomia”) są w pracy *explicite* problematyzowane i objaśniane, co świadczy o wysokiej świadomości metodologicznej doktoranta.

Styl narracji można określić jako analityczno-argumentacyjny, pozbawiony zbędnej emfazy czy publicystycznych uproszczeń. Doktorant unika potoczności oraz mętnych metafor, a jednocześnie zachowuje czytelność i przejrzystość wyводу, co jest szczególnie istotne w pracy obejmującej zagadnienia interdyscyplinarne. Struktura zdań jest zrównoważona, a argumentacja rozwijana jest stopniowo, z wyraźnym rozróżnieniem między rekonstrukcją cudzych stanowisk a ich autorską oceną.

Pod względem warsztatowym rozprawa prezentuje bardzo wysoki poziom naukowy. Doktorant wykazuje się doskonałą orientacją w literaturze przedmiotu,

obejmującej zarówno klasyczne teksty filozoficzne, jak i najnowsze publikacje z zakresu etyki sztucznej inteligencji, etyki maszyn oraz badań nad sztucznymi agentami moralnymi. Dobór źródeł jest trafny, aktualny i funkcjonalnie podporządkowany problematyce pracy. Co istotne, autor nie poprzestaje jedynie na referowaniu stanowisk, lecz poddaje je krytycznej analizie, ujawniając ich założenia filozoficzne, ograniczenia oraz konsekwencje.

Aparat naukowy pracy został opracowany z należytą starannością. Przypisy są używane w sposób adekwatny, nienadmierny, a jednocześnie wystarczający do precyzyjnego dokumentowania źródeł i prowadzenia twórczego dialogu z istniejącą literaturą naukową w dyskutowanym przedmiocie. Bibliografia jest pokaźna, uporządkowana i obejmująca kluczowe pozycje z omawianego obszaru badawczego, w tym prace autorów uznawanych za liderów współczesnej debaty naukowej nad etyką sztucznej inteligencji (m.in. Nick Bostrom, Stuart Jonathan Russell, Luciano Floridi, Mark Coeckelbergh, Joanna Joy Bryson, Kate Crawford).

Na uznanie zasługuje również popisowa umiejętność doktoranta w zakresie krytycznej syntezy ogromnie rozpiętego i zróżnicowanego materiału badawczego. Mgr Mróz nie gubi się w rozległej literaturze przedmiotu ani nie popada w encyklopedyczność, lecz zachowuje wyraźną linię problemową, podporządkowując analizowane koncepcje zasadniczemu pytaniu o możliwość i sens przypisywania sztucznej inteligencji, w tym zwłaszcza sztucznym sieciom neuronowym, moralnej sprawczości. Świadczy to niewątpliwie o jego wysokiej kulturze metodologicznej, biegłości i fachowości w podejmowanej problematyce oraz nieczęstej wśród młodych badaczy dojrzałości naukowej.

Podsumowując, język dysertacji oraz jej walory warsztatowe należy ocenić bardzo wysoko. Rozprawa spełnia – a miejscami wyraźnie przekracza – standardy stawiane pracom doktorskim w dyscyplinie nauk teologicznych i filozoficznych. Doktorant wykazuje zaawansowaną kompetencję w obszarze prowadzenia samodzielnych badań naukowych, posługiwania się aparatem pojęciowym oraz krytycznej pracy z literaturą przedmiotu.

Merytoryczna wartość rozprawy

Sama w sobie oraz na tle licznych znanych mi rozpraw doktorskich recenzowana praca przedstawia szczególnie wysoką wartość merytoryczną. Przejawia się ona na kilku poziomach. Po pierwsze, praca stanowi bodaj pierwsze tego rodzaju, kompetentne i pogłębione opracowanie jednego z kluczowych problemów współczesnej etyki sztucznej inteligencji, jakim jest pytanie o możliwość, sens i konsekwencje przypisywania

systemom AI statusu moralnych agentów. Doktorant wykazuje ekspercką znajomość aktualnego stanu badań oraz głównych osi sporów, toczących się w tej dziedzinie.

Po drugie, rozprawa mgr. Macieja Mroza wnosi istotny wkład do badań naukowych nad etyką sztucznej inteligencji, szczególnie poprzez systematyczną i rzetelnie uargumentowaną krytykę paradygmatu *value alignment* (dostosowania sztucznej inteligencji do ludzkich wartości i celów) oraz związanego z nim problemu kontroli sztucznej inteligencji. Doktorant starannie pokazuje, że wiele propozycji obecnych w literaturze przedmiotu opiera się na uproszczonym rozumieniu moralności jako zbioru preferencji lub funkcji celu, co prowadzi do powstania nieadekwatnych i redukcjonistycznych modeli normatywnych AI. Krytyka ta jest prowadzona w sposób wyważony, oparty na rzetelnej rekonstrukcji stanowisk, a nie na arbitralnej polemice o charakterze eseistycznym.

Na szczególne uznanie zasługuje interdyscyplinarna kompetencja doktoranta, który z niezwykłą swobodą i biegłością porusza się zarówno po literaturze filozoficznej, jak i po pracach publikowanych przez badaczy AI. Rozprawa trafnie identyfikuje i analizuje napięcia wynikające z różnic metodologicznych, ontologicznych i teoriopoznawczych między tymi obszarami, wskazując je jako jedno z głównych źródeł nieporozumień w debacie o moralności sztucznej inteligencji.

Z punktu widzenia nauk teologicznych istotny jest również antropologiczny reżim wywodu. Mgr Mróz konsekwentnie broni tezy, że refleksja nad sztucznymi agentami moralnymi nieuchronnie prowadzi do pogłębionych pytań o naturę ludzkiej moralności, podmiotowości, wolności, sprawczości i odpowiedzialności. Tym samym doktorant nie tylko analizuje AI jako obiekt badań, lecz także – co zasługuje, moim zdaniem, na szczególne uwypuklenie – wykorzystuje te analizy do ukazania sztucznej inteligencji jako swego rodzaju zwierciadła, pozwalającego lepiej uchwycić specyfikę ludzkiego działania moralnego.

Oryginalność ujęcia problemu badawczego

Oryginalność rozprawy Macieja Mroza polega przede wszystkim na sposobie postawienia i przepracowania problemu moralności sztucznych agentów, a nie na próbie tworzenia nowej, konkurencyjnej teorii normatywnej, zmętniającej dyskurs, niewiele wnoszącej do toczonych debat oraz tworzącej sztuczny tłok w i tak już gęstym tłumie tego rodzaju akademickich ujęć. Doktorant świadomie unika ambicji systemotwórczych, koncentrując się na rzeczowej, krytycznej analizie założeń filozoficznych, leżących u podstaw współczesnych projektów moralności obliczeniowej i sztucznych agentów

moralnych. Takie podejście należy uznać za dojrzałe metodologicznie i dobrze uzasadnione w kontekście aktualnego stanu badań.

Nowatorski charakter pracy ujawnia się szczególnie w konsekwentnym odslanianiu redukcjonistycznych przesunięć pojęciowych, które zachodzą, gdy klasyczne pojęcia filozofii i etyki (takie, jak podmiotowość, autonomia, sprawczość, intencjonalność, wartość czy odpowiedzialność,) przenoszone są na grunt inżynierii algorytmicznej. Doktorant nie ogranicza się do krytyki poszczególnych rozwiązań technicznych, lecz pokazuje, że wiele problemów etycznych, przypisywanych „niedojrzałości”, niebiologiczności, bezosobowości czy schematyczności działania technologii sztucznej inteligencji, ma w istocie charakter błędu kategoryjnego – tu przoduje błąd antropomorfizacji AI – lub nieuprawnionego, wszak redukcjonistycznego funkcjonalizmu.

Istotnym elementem oryginalności ocenianej pracy jest również połączenie w niej debaty nad sztucznymi agentami moralnymi z refleksją nad kreatywnym charakterem czynu moralnego. Wprowadzenie kategorii twórczości moralnej jako argumentu przeciwko stanowiskom opowiadającym się za tezą o komputacyjnym charakterze moralności, stanowi autorski wkład mgr. Mroza w toczoną obecnie dyskusję i wyraźnie odróżnia jego rozprawę od dominujących ujęć analitycznych i inżynierskich. Ten wątek, rozwijany szczególnie w rozdziale czwartym, poszerza horyzont debaty i wnosi do niej perspektywę rzadko obecną w literaturze dotyczącej etyki sztucznej inteligencji.

Na uwagę zasługuje również osadzenie poruszanych w pracy zagadnień i analiz w perspektywie teologicznej, przy jednoczesnym zachowaniu spójności z dyskursem filozoficznym i inżynierskim. Co ważne, doktorant nie instrumentalizuje teologii i nie sprowadza jej – by tak rzec – do roli magazynu gotowych odpowiedzi moralnych, lecz traktuje ją jako pogłębione zaplecze antropologiczne, umożliwiające zobiektywizowaną, krytyczną ocenę ambicji przypisywania maszynom moralnej sprawczości. Takie ujęcie zdecydowanie wyróżnia się na tle dominującej, sekularnej literatury przedmiotu.

Uwagi krytyczne

Pomimo niekwestionowanej, obiektywnie wysokiej merytorycznej wartości rozprawy, zgodnie z rolą, jaka przypada do wypełnienia recenzentowi, chciałbym wskazać również kilka kwestii, które mogą budzić pytania lub wątpliwości albo stanowić punkt wyjścia do koniecznego doprecyzowania dominującej w pracy linii argumentacyjnej.

Po pierwsze, silna koncentracja doktoranta na krytycznym wymiarze etyki maszyn i paradygmatu sztucznych agentów moralnych sprawia, że alternatywne propozycje i rozwiązania pojawiają się w pracy raczej w sposób implicytny niż systematyczny.

Doktorant przekonująco wykazuje ograniczenia funkcjonalistycznych i obliczeniowych ujęć moralności, jednak nie rozwija w pełni pozytywnego modelu interpretacyjnego, który mógłby stanowić konkurencyjną ramę dla refleksji nad projektowaniem systemów AI. Choć taki wybór jest metodologicznie uzasadniony i zgodny z deklarowanym celem pracy, powstaje pytanie, czy przynajmniej szkic bardziej wyraźnej alternatywy normatywnej nie wzmocniłby siły oddziaływania formułowanych w dysertacji wniosków.

Po drugie, perspektywa teologiczna, mimo że wyraźnie zaakcentowana i ważna z punktu widzenia realizacji celów rozprawy, pozostaje miejscami raczej horyzontem interpretacyjnym niż przedmiotem bezpośredniej, pogłębionej analizy. Doktorant świadomie unika teologicznej normatywizacji naukowej debaty nad sztuczną inteligencją, co poczytuję akurat za zaletę pracy, jednak może to rodzić wrażenie, że potencjał teologii moralnej – zwłaszcza w zakresie teorii sumienia, odpowiedzialności i rozeznania – nie został w pełni wykorzystany. Problem sumienia nie został *explicite* podniesiony w pracy, tak jakby działania sumieniowe nie były uznawane przez doktoranta za istotny i niezbywalny składnik oraz konstytutywny czynnik działania moralnego. Rozważenie kwestii możliwości wygenerowania algorytmicznego sumienia w modelach AI wydaje mi się równie ważnym problemem badawczym, jak te, które porusza w swojej rozprawie doktorant. Mimo to zagadnienie sumienia pozostaje w jego pracy milczącym nieobecnym. A przecież istnieją w tej materii nieliczne, co prawda, acz ważne i interesujące publikacje, jak np. artykuł naukowy Guntera Meissnera pt. *Artificial Intelligence: Consciousness and Conscience*¹ czy Krzysztofa Wieczorka pt. *The Conscience of a Machine? Artificial Intelligence and the Problem of Moral Responsibility*². W artykule z 2025 r. prof. Agnieszka Lekka-Kowalik pisze w tej sprawie: „[...] jeśli AI ma symulować ludzkie myślenie i zachowanie, to należy ją wyposażyć w sztuczne sumienie – zdolność sądzenia o własnych (i cudzych) czynach w aspekcie dobra i zła moralnego”³.

Potencjalny czytelnik ocenianej rozprawy może odczuwać poznawczy niedosyt w związku z tymi miejscami pracy, w których przydałoby się bardziej bezpośrednio skonfrontowanie stanowisk filozofii AI z konkretnymi nurtami teologii moralnej.

¹ Zob. Gunter Meissner, „Artificial Intelligence: Consciousness and Conscience”, *AI & Society*, 2020, nr 35, s. 225–235.

² Zob. Krzysztof Wieczorek, „The Conscience of a Machine? Artificial Intelligence and the Problem of Moral Responsibility”, *Er(r)go. Teoria–Literatura–Kultura*, 2021, nr 42(1), s. 15–34.

³ Agnieszka Lekka-Kowalik, „W cieniu rozkwitających AI”, [w:] *Człowiek wobec sztucznej inteligencji. Dwugłos nauki*, red. Jerzy Kaczorowski, Poznań: Polska Akademia Nauk, Oddział w Poznaniu, 2025, s. 53. Por. A. Lekka-Kowalik, *Czy AI ma sztuczne sumienie?*, artykuł zamieszczony na stronie internetowej Katolickiego Uniwersytetu Lubelskiego im. Jana Pawła II w Lublinie, dział „Kultura”, zakładka „Aktualności / Wydarzenia”. Dostęp on-line: https://www.kul.pl/ekspert-kul-czy-ai-ma-sztuczne-sumienie,art_112253.html.

Po trzecie, rozprawa mgr. Mroza koncentruje się głównie na zachodnim dyskursie filozoficznym i technologicznym. Choć wybór ten jest zrozumiały i uzasadniony zakresem problemowym rozprawy, to jednocześnie ogranicza szerszą perspektywę kulturową. W kontekście globalnego rozwoju sztucznej inteligencji można postawić pytanie, na ile wnioski dotyczące antropomorfizacji, sprawczości moralnej czy kreatywności czynu moralnego zachowują swoją uniwersalność w odmiennych tradycjach filozoficznych i religijnych.

Wreszcie, z punktu widzenia odbiorcy zainteresowanego praktycznymi implikacjami dysertacji, konsekwencje analiz dla konkretnych zastosowań AI (np. w nauce, edukacji, sztuce, medycynie, wojskowości, przemyśle, handlu etc.) pozostają w rozprawie jedynie wstępnie zarysowane niż analitycznie rozwinięte.

Postulaty dalszych badań

Wskazane powyżej uwagi bynajmniej nie umniejszają naukowej wartości rozprawy, lecz raczej otwierają naturalnie się wyłaniające i obiecujące kierunki dalszych badań, do których autor ocenianej pracy wydaje się szczególnie predestynowany.

Po pierwsze, zasadne byłoby rozwinięcie pozytywnej koncepcji relacji między moralnością ludzką a systemami AI, która nie redukowałaby moralności do algorytmu, a jednocześnie pozwalała sensownie opisać wspomagającą rolę systemów technicznych w ludzkich procesach decyzyjnych w sytuacjach moralnie złożonych. Mogłoby to przyjąć formę teorii „moralnego wsparcia” (*moral decision support*), wyraźnie odróżnionej od idei autonomicznych moralnych agentów.

Po drugie, wartościowym kierunkiem dalszych badań byłoby pogłębienie teologicznej antropologii technologii, zwłaszcza w dialogu z klasycznymi i współczesnymi teoriami sumienia, cnoty i odpowiedzialności. Szczególnie interesujące wydaje mi się tu rozwinięcie wątku kreatywnego charakteru czynu moralnego w odniesieniu do teologii działania ludzkiego oraz chrześcijańskiej koncepcji wolności.

Po trzecie, dalsze badania mogłyby objąć porównawczą analizę międzykulturową podejść do moralności maszyn, uwzględniając tradycje filozoficzne i religijne spoza kręgu zachodniego (np. konfucjanizm, buddyzm czy etykę islamską). Taka perspektywa mogłaby zarówno potwierdzić, jak i twórczo zakwestionować niektóre założenia obecne w dominującym zachodnim dyskursie etyki sztucznej inteligencji.

Po czwarte, interesującym rozwinięciem przedstawionych w dysertacji analiz byłoby zastosowanie wypracowanych w niej narzędzi pojęciowych do studiów przypadków konkretnych systemów AI, zwłaszcza tych, które już dziś funkcjonują w obszarach wrażliwych moralnie. Pozwoliłoby to sprawdzić praktyczną nośność krytyki

pojęciowej i wzmocnić dialog między refleksją filozoficzną i teologiczną a projektowaniem technologii.

Chciałbym podkreślić, że przedstawione uwagi krytyczne mają jedynie charakter uzupełniający i perspektywiczny, a nie kardynalny. Nie podważają one bynajmniej ani spójności wywodu, ani zasadniczej wartości naukowej rozprawy, lecz wskazują na jej rozwojowy potencjał badawczy. Dysertacja mgr. Macieja Mroza stanowi solidną podstawę do dalszych, pogłębionych badań naukowych oraz świadczy o jego gotowości do podjęcia pracy naukowej.

Konkluzja

W świetle przedłożonych opinii z pełnym przekonaniem stwierdzam, że rozprawa doktorska Pana mgr. Macieja MROZA z nadstatkiem spełnia warunki określone w art. 187 Ustawy z dnia 20 lipca 2018 r. *Prawo o szkolnictwie wyższym i nauce* (Dz. U. z 2018 r., poz. 1668, z późn. zm.). W związku z tym kieruję do Rady Naukowej Dyscypliny nauki teologiczne Uniwersytetu im. Adama Mickiewicza w Poznaniu **wniosek o dopuszczenie Doktoranta do dalszych, przewidzianych prawem, etapów przewodu doktorskiego**.

Pragnę również niniejszym przedstawić Radzie Naukowej Dyscypliny nauki teologiczne Uniwersytetu im. Adama Mickiewicza w Poznaniu mój **wniosek o wyróżnienie rozprawy doktorskiej mgr. Macieja Mroza**. Oto uzasadnienie mojego wniosku: rozprawa doktorska mgr. Macieja Mroza zasługuje na wyróżnienie ze względu na jej wyjątkowo wysoki poziom merytoryczny, metodologiczny i warsztatowy, a także oryginalność ujęcia podjętego w niej problemu badawczego, który należy do jednych z najważniejszych i najbardziej aktualnych zagadnień naukowych oraz cywilizacyjnych wyzwań współczesności.

Praca doktorska mgr. Mroza podejmuje problem moralności sztucznych agentów oraz etycznych implikacji rozwoju sztucznej inteligencji, osadzając go w kontekście filozoficznym, etycznym i teologicznym. Doktorant nie ogranicza się do referowania wypracowanych na gruncie interdyscyplinarnych badań nad sztuczną inteligencją stanowisk, lecz prowadzi samodzielną, wysoce kompetentną, systematyczną i krytyczną analizę założeń, leżących u podstaw dominującego w literaturze naukowej paradygmatu sztucznych agentów moralnych, ujawniając jego ograniczenia, przesunięcia pojęciowe oraz konsekwencje antropologiczne i normatywne. Tego rodzaju krytyka paradygmatyczna, jaką przeprowadza w dysertacji doktorant, należy do rzadkości w wśród publikacji z tego obszaru, zdominowanych przez redukcjonistyczne podejścia funkcjonalistyczne i inżynieryjne.

Na szczególne podkreślenie zasługuje podkreślana przeze mnie w recenzji dojrzałość metodologiczna doktoranta, który z ogromnym znawstwem i biegłością porusza się na styku filozofii, etyki technologii, teologii oraz współczesnych badań nad sztuczną inteligencją. Mgr Mróz wykazuje doskonałą orientację w międzynarodowej literaturze przedmiotu, obejmującej zarówno klasyczne teksty filozoficzne, jak i najnowsze publikacje z zakresu etyki sztucznej inteligencji oraz etyki maszyn. Analiza ta nie ma charakteru kompilacyjnego, lecz prowadzi do spójnych i dobrze uzasadnionych wniosków.

Wyróżniającą cechą rozprawy jest również konsekwentne unikanie redukcjonizmu w opisie moralności sztucznych agentów. Doktorant przekonująco argumentuje, że moralność nie daje się sprowadzić do algorytmu, funkcji celu ani procedury decyzyjnej, a próby takiej formalizacji prowadzą do błędów kategoryalnych i nieuprawnionej antropomorfizacji systemów technicznych. Szczególnie wartościowy jest w tym kontekście wątek twórczego i osobowego charakteru czynu moralnego, rzadko obecny w dyskusjach nad etyką AI, a rozwinięty w pracy w sposób oryginalny i pogłębiony.

Na uwagę zasługuje także umiejętne i wyważone wykorzystanie perspektywy teologicznej, która nie pełni w pracy funkcji normatywnego uproszczenia ani zewnętrznego autorytetu, lecz stanowi zaplecze antropologiczne i hermeneutyczne, umożliwiające krytyczną ocenę współczesnych koncepcji moralności maszyn. Doktorant zachowuje przy tym pełną kompatybilność z pozateologicznym dyskursem naukowym, co czyni rozprawę wartościową również dla badaczy spoza teologii.

Pod względem językowym i edytorskim rozprawa mgr. Macieja Mroza spełnia najwyższe standardy pracy doktorskiej. Została napisana w języku angielskim na poziomie odpowiadającym międzynarodowym normom akademickim, z zachowaniem precyzji terminologicznej, klarowności wywodu oraz wysokiej kultury argumentacyjnej. Aparat naukowy, struktura pracy oraz sposób prowadzenia polemiki z ustaleniami zmaterializowanymi w publikacjach uznanych badaczy sztucznej inteligencji z całego świata świadczą o zaawansowanych kompetencjach naukowych doktoranta.

Sformułowane uwagi, spostrzeżenia i oceny pozwalają stwierdzić, że rozprawa doktorska mgr. Macieja Mroza nie tylko spełnia wymagania stawiane zgodnie z prawem i uzusem akademickim pracom doktorskim, lecz wyraźnie ponad nie wykracza, wnosząc trwałe i oryginalne wkłady w rozwój badań nad etyką sztucznej inteligencji, filozofią techniki oraz teologią. Z tych względów, w moim najgłębszym przekonaniu, dysertacja mgr. Mroza w pełni zasługuje na wyróżnienie.

Jednocześnie **rekomenduję tę pracę do publikacji w formie monografii naukowej**, zarówno w języku angielskim, jak i polskim. Ukazanie się takiej pracy, zwłaszcza w języku polskim, z pewnością wypełni dotkliwą lukę i znacznie wzbogaci

międzynarodowe i lokalne repozytorium naukowe w obszarze badań nad sztuczną inteligencją.

Recenzję sporządziłem: 27.01.2026.

.....
podpis recenzenta