Prof. Dr hab. Ewa Agnieszka Lekka-Kowalik
Wydział Filozofii
Katolicki Uniwersytet Lubelski Jana Pawła II

Recenzja pracy doktorskiej mgr Macieja Mroza pt. Morality of Artificial Agents (Moralność sztucznych agentów)

Poznań 2025, ss. 155 + 16 s. bibliografia

Przedstawiona do recenzji praca doktorska napisana jest w języku angielskim, a jej treść odpowiada tytułowi pracy. Składa się ze spisu treści, wstępu, czterech rozdziałów, zakończenia oraz obszernej bibliografii. Nie ma streszczenia w języku polskim, co wydaje się być formalnym wymogiem prac pisanych w obcych językach.

Wstęp zawiera wszystkie metodologicznie niezbędne elementy. Doktorant w badaniach sztucznych agentów zasadniczo odwołuje się – słusznie – do filozofów i do teoretyków AI. Natomiast za istotną cechę dysertacji uznaje zakorzenienie jej w teologii. Jest to zrozumiałe, skoro praca pisana jest na Wydziale Teologicznym i doktorat ma być z dyscypliny zwanej „nauki teologiczne". Owo zakorzenienie odwołuje się do faktu, iż ludzkie działania i jego wytwory są „miejscami teologicznymi", w których manifestuje się natura i przeznaczenie człowieka. Słusznie też Doktorant podkreśla, że rozwój AI stanowi dziś istotny kontekst, w którym pytania o człowieka muszą być stawiane. Takie rozumienie teologii ks. prof. Stanisław Kamiński, przedstawiciel lubelskiej szkoły filozoficznej, nazywał „rewelacjonizacją świata". Bardzo szkoda, że Pan Mróz nie wrócił do tego wątku i nie pokazał w szczegółach – po wszystkich rozważaniach – owych sztucznych „podmiotów" jako „teologicznego miejsca". Jest to brak, który warto uzupełnić choćby podczas publicznej obrony. Podana na stronie tytułowej polska wersja tytułu używa terminu „agent". Termin ten faktycznie rozpanoszył się w języku polskim, choć faktycznie chodzi o „podmiot działający". Problematyczność użycia terminu „podmiot" w kontekście AI jest oczywista, co zresztą jest rozważane w dysertacji. Język angielski jest tu wygodniejszy, ponieważ termin „agent" nie niesie z sobą tych wszystkich konotacji, które niesie polski termin „podmiot". Kate Darling z MIT zaproponowała termin „aktant", by zaznaczyć moc działania AI, a odróżnić od ludzi (zob. Darling, K. (2016). Extending Legal Protection to Social Robots. In R. Calo, M. Froomkin, I. Kerr (eds.). *Robot Law*, Edward Elgar Publishing, Cheltenham, UK.).

Przedmiotem rozważań Doktoranta jest sztuczna inteligencja widziana właśnie jako aktant, na który narzuca się warunek działania w sposób moralny i/lub bycia moralnym. Zwrot „i/lub" sygnalizuje kwestię doskonale rozważoną w dysertacji: „moralne zombi" versus „moralny podmiot", do czego później nawiążę. Autor stawia swym rozważaniom trzy cele: (1) przedstawienie debaty nad możliwością, racjami i warunkami zbudowania moralnego aktanta; (2) przeanalizowanie kontekstów, w których ta debata się rodzi; (3) wskazanie konsekwencji zasadniczych stanowisk w tej debacie (por. s. 9). Praca ma więc charakter metaprzedmiotowy. Uważam, że cele zostały zrealizowane. Słusznie Autor podkreśla dynamiczną sytuację, w której prowadzi swe analizy – AI bardzo szybko się rozwija i jest zjawiskiem wielowymiarowym. Stąd

pewne pytania – jak np. o status moralny AI – muszą być z góry wykluczone z rozważań.

Rozdział 1 szkicuje kontekst, w którym rodzi się pytanie o moralność aktantów. Punkt 1 tego rozdziału pokazuje problemy z definiowaniem AI oraz ich typologią, szkicuje – ważną dla dalszych analiz – kwestię antropomorfizowania AI oraz główne obszary rozważań należące do etyki AI, czyniąc przy tym szereg rozróżnień pojęciowych (np. odpowiedzialna AI i AI godna zaufania). Słusznie Autor podkreśla, że AI nie jest kwestią li tylko techniczną, ale także społeczną i etyczną (a zapewne także polityczną i ekonomiczną). Bardzo słusznie też podkreśla, że dyskutowane kwestie rodzą się z samego istnienia i funkcjonowania AI, a nie z ich niemoralnego wykorzystywania (zob. s. 30). Rodzi to kwestię ogólniejszą: AI jest znakomitym przykładem nasycenia artefaktów wartościami (zob. Lizut, R.A., *Technika a wartości. Spór o aksjologiczną neutralność artefaktów*, Lublin 2014). Owo nasycenie artefaktów wartościami sprawia, że nie tylko ich wykorzystanie, ale także budowanie i rozpowszechnianie staje się kwestią etyczną. Autor nie podejmuje tematu nasycenia techniki wartościami, choć w rozdziale 3 stawia pytanie, czy powinniśmy budować moralne aktanty (zob. 3.2), co jest naturalną konsekwencją uznania nasycenia AI wartościami. Warto by ten wątek rozwinąć. Stawia natomiast inną istotną kwestię, która nazywa „bilateralnym zanieczyszczeniem" (*bilateral contamination*): badania AI przyjmują antropomorfizujący opis procesów komputacyjnych, a nauki kognitywne adoptują komputacyjne ujęcia ludzkich procesów poznawczych (zob. s. 21n). Rozważania bardzo dobrze pokazują, że taka kontaminacja nie jest obojętna dla wyników badawczych i rozumienia człowieka – może nieść z sobą zniekształcanie rzeczywistości. Podobne analizy znajdziemy w watykańskim dokumencie *Antiqua et Nova (*do którego Pan Mróz sięga), a dotyczą terminu „inteligencja" bezrefleksyjnie stosowanym do człowieka i do techniki. Sam Autor nie zajął się – być może szkoda – problemami z rozumieniem terminu „inteligencja". Rozdział 1 jest bardzo informatywny, przywołuje ważnych autorów i ważne dokumenty. Wyraźnym niedopatrzeniem jest natomiast brak odwołania do tzw. *Asilomar AI Principles* z 2017 (https://futureoflife.org/open-letter/ai-principles/), w którym to dokumencie (podpisanym przez czołowe postaci i organizacje świata AI) pojawiają się *explicite* problemy i pojęcia istotne dla analiz zawartych w pracy. Natomiast niewątpliwie rozdział 1. spełnia swój cel: naszkicowanie kontekstu dalszych rozważań, i z korzyścią mogłyby go czytać wszystkie osoby zainteresowane AI. Dobrze też pokazuje kluczowy problem etyki AI: „przetłumaczenie" etycznych norm na techniczne procesy (zob. 1.3.4 i doskonały tytuł tego punktu).

Rozdział 2. jest kluczowy dla rozważań, bowiem skupiony jest na tytułowym pojęciu *artificial agency* i jego filozoficznych fundamentach, poczynając od rozumienia podmiotowości. Uwzględnia też paradygmatyczny zwrot w rozważaniach: zamiast rozważać sztuczną inteligencję mamy rozważać sztucznego agenta (podmiot działający), niekoniecznie posiadającego ludzkie podmiotowe wyposażenie (2.2). W kolejnych punktach dobrze jest pokazane, że uznajemy pewne sztuczne twory, takie jak instytucje czy korporacje, za „podmioty" działające i wobec tego pytanie brzmi nie: czy istnieją sztuczne podmioty działające?, ale: czy można do nich zaliczyć AI?. Rozważania nad formami sztucznych podmiotów (2.3) oraz model kolektywnego i rozproszonego działania (2.4) prowadzą do istotnego wniosku: granica między naturalnym i sztucznym podmiotem działającym wydaje się zanikać (s. 52). Choć Autor tego nie podkreśla, rozważania te uzasadniają postawienie pytania o AI jako agenta własnie – i jest to kwestia faktyczna, a nie czysto akademicki problem. Swe analizy „paradygmatycznego przesunięcia" Doktorant uzupełnia analizami historycznymi

rozumienia działania i podmiotu działania (2.5; od Arystotelesa do współczesnych autorów). Ów historyzm (obecnym i w innych rozdziałach) jest dobrym metodologicznym zabiegiem, ponieważ pokazuje, że „stare" narzędzie pojęciowe i ideowe mogą być przydatne do ujaśnienia i rozwiązania współczesnych kwestii. Autor nie ulega więc modnemu dziś „byciu na czasie", co uważam za istotny walor pracy. Używam tu terminu „historyzm" charakteryzującego lubelską szkołę filozoficzną, ale uważam, że doskonale pasuje do podejścia Pana Mroza: chodzi o uwzględnienie w badaniach filozoficznych historii wyjaśnianych problemów, głównie pod kątem kontekstu ich powstania i rozwoju, pozwalającego odkryć w ramach określonych systemów filozoficznych, mimo pojęciowego zróżnicowania, stałe aspekty problemów oraz wpływ na ich interpretacje przyjmowanych założeń i metod badawczych, jak też sposobów wyjaśniania i uzasadniania formułowanych tez (zob. A. Lekka-Kowalik & T. Duma, "Via ad veritatem. O metodach uprawiania filozofii w szkole lubelskiej", w: A. Lekka-Kowalik, P. Gondek (red.), *Lubelska Szkoła Filozoficzna. Historia – koncepcje – spory,* Wydawnictwo KUL, Lublin 2019, 147-173). Historyczne rozważania pozwalają Autorowi wskazać konieczne atrybuty bytu, który można uznać za działający (2.7) oraz wskazuje, że owo bycie podmiotem działającym jest stopniowalne (2.7.1). Punkt 2.8 jest kluczowy dla pytania o konsekwencje uznania AI za podmiot działający, ponieważ wprowadza pojęcie funkcjonalnego podmiotu działającego (opis obejmuje działania, a nie metafizyczne wyposażenie podmiotu) oraz kwestię realizowania własności funkcjonalnych w rozmaitych substratach (*multiple realizability*). Współczesne ujęcia pokazują, że owa *multiple realizability,* a także niezawodności i solidność działania (*reliability*) są sprawą stopnia: sztuczne systemy mogą realizować pewne aspekty działającego podmiotu, a innych nie, mogą być wysoce solidne w jednych aspektach (np. czasowym, ekologicznym), a w innych nie. Autor bardzo dobrze referuje dyskusję, ale szkoda, że nie pokazał konsekwencji (czy nie naszkicował problemów) prawnych i moralnych takiego ujęcia sztucznego podmiotu działającego. A przecież w tym przypadku stajemy przed problemem wyboru, które aspekty są na tyle istotne, że AI ich nie respektujący nie może/nie powinien być dopuszczony do działania. Bardzo krótko (p. 2.9) Pan Mróz stawia problem sztucznej osoby i jej relacji do działania (pół s. 64), odróżnia osobę prawną i moralną oraz wprowadza pojęcia „osoby elektronicznej" (kolejne pół strony) oraz wskazuje na cierpliwość i podatność na działania innych jako istotne cechy podmiotu działającego. Szkoda, że nie rozbudował tych punktów, bo tam pojawiają się kluczowe problemy: podobieństwo działań AI do działań osób ludzkich (kwestia świadomości), rozumienie osoby ludzkiej i konsekwencje rozszerzania zakresu tego pojęcia (a więc i zmiana treści), co rodzi kwestie ontologiczne. Autor wspomina o rozmaitych kwestiach z tego obszaru, ale nie jest to analiza, a raczej deklaracja, i to raczej sprawozdanie z czyichś poglądów. A szkoda, bo tu do pewnego stopnia rozstrzygają się niektóre kwestie istotne dla rozdziału 3.

Rozdział 3. podejmuje kwestię sztucznych podmiotów moralnych. Autor jasno i precyzyjnie stawia tę kwestię: działania AI mogą być oceniane jako moralnie dobre/słuszne albo moralnie złe/niesłuszne. Czy wobec tego samo działające AI może być moralne/niemoralne? Doktorant najpierw pokazuje jak można rozumieć etyczność bytów działających (idzie to za J. Moorem; 3.1.1), a następnie podejścia do tworzenia etycznych maszyn (3.1.2) oraz tzw. moralny test Turinga (3.1.3). Słusznie podkreśla, że oprócz wyboru podejścia „od dołu", „od góry" czy hybrydowego, jest także problem wyboru etyki jako teorii normatywnej, a nawet normatywności pośredniej (przedstawione jest podejście N. Bostroma). Już tu zaznacza też

wagę problemu: maszyny są postrzegane jako bardziej neutralne w swych sądach i decyzjach niż ludzie (s. 75). Nie rozbudował natomiast konsekwencji faktu, że AI jest widziana jako moralnie lepsza niż my, ludzie, a są one i społecznie, i moralnie istotne (zob. A. Lekka-Kowalik, *Człowiek wobec sztucznej inteligencji,* red. Jerzy Kaczorowski, Poznań 2025, s. 49-58.). Rozdział jest bardzo erudycyjny i stanowi doskonałe źródła do wszystkich, którzy chcą się zapoznać z tzw. etyką maszyn (*machine ethics*) i jej głównymi problemami. Sam Doktorant nie zabiera wyraźnie głosu w dyskutowanych kwestiach, raczej analitycznie streszcza rozmaite podejścia. Punkt 3.2 (s. 81n) jest bodaj najistotniejszy, ponieważ stawia kwestię normatywną: czy *powinniśmy* dążyć do zbudowania moralnych maszyn? Słusznie Autor podkreśla, że pozornie jasna odpowiedź: skoro AI decyduje i działa, a rezultaty są moralnie istotne, to powinniśmy wyposażyć ją w jakiś rodzaj moralnego rozumowania, wcale nie jest oczywista. W kolejnych punktach przedstawia argumenty za i przeciw temu przedsięwzięciu (jak sądzę w numeracji jest błąd, bo dwa razy występuje 3.2.1). Jest to bardzo dobra prezentacja, oparta na najnowszej literaturze, nawet jeśli argumenty „nie da się tego zrobić" i „nie powinno się tego zrobić" nie są wyraźnie oddzielone. Rozdział kończy się kwestią tzw. ujednolicenia wartości (*value alignment*), postawionej już w Zasadach z Asilomar, na które wskazywałam jako brak w literaturze cytowanej: „Wysoce autonomiczne systemy sztucznej inteligencji powinny być projektowane w taki sposób, aby ich cele i zachowania były zgodne z wartościami ludzkimi przez cały okres ich działania" (Zasady z Asilomar, p. 10; https://instytutsprawobywatelskich.pl/sztuczna-inteligencja-zasady-z-asilomar/), co słusznie jest pokazane jako prawdziwe wyzwanie dla konstrukcji AI i to nie tylko dla superinteligencji. Dla analiz problemu ujednolicenia wartości zaproponowany jest model agent-pryncypał (s. 93) z kilkoma uwagami. Jest to bardzo interesująca propozycja, oparta na znanym modelu, gdzie pryncypał deleguje zadania do wykonania przez agenta, co rodzi rozmaite trudności. Nie jest jasne, czy to jest pomysł od kogoś wzięty (nie ma żadnego przypisu, choć jest fragment, który wygląda jak cytat) czy też to własny pomysł autora. W każdym razie warto byłoby to podejście rozwinąć, w jakim bowiem sensie zachowanie AI wykonującej zadanie może być „niespodziewane" czy inne niż to zaprogramowane? A do takiej rozbieżności odwołuje się Autor (p. 93). Problem kontroli działań AI (3.3.2), tworzenia systemów „uwartościowanych" (3.3.3), etyczna krytyka postulatu ujednolicenia wartości (3.3.4) oraz dowody na niepowodzenie ujednolicania wartości w przypadku dużych modeli językowych (LLM) (3.3.5) doskonale pokazują, z jakim problemem przychodzi się nam mierzyć. I jasno widać, że nie jest to problem teoretyczny, bowiem trwają intensywne badania nad realizacją moralnych maszyn. Jest to niewątpliwie zaleta rozdziału: ma nie tylko duży ładunek erudycyjny, ale sięga do faktycznych rezultatów w dziedzinie AI.

Rozdział 4. ma na celu pogłębienie kluczowych wątków zarysowanych w poprzednich rozdziałach, a dokładniej przebadanie centralnych wymiarów moralności i obecności tych wymiarów w działaniach sztucznych podmiotów (s. 103). Ostatecznie pozwoliłoby to ocenić zasadność przypisywania AI moralności. To stąd tytuł rozdziału w formie pytania: czy moralność ma charakter komputacyjny? Próby odpowiedzi na to pytanie – twierdzi Doktorant – próbują łączyć dwie idee: moralność w sensie fenomenalnym, gdy podmiot moralny ma mieć świadomość czy doświadczenie jakości oraz moralność w sensie funkcjonalnym, gdy moralność jest rozumiana jedynie jako cecha zachowania. Daje przykłady takiego połączenia (4.1): „przetłumaczenie" teologicznej ontologii na funkcjonalny układ czy idea wychowywania

robotów oraz przeprowadza ich dobrą, ale dość skrótową krytykę. Natomiast słusznie podkreśla – i jego rozważania dobrze to ilustrują – że debaty nad AI potrzebują nie tylko technicznych, ale i filozoficznych fundamentów (s. 108). Kolejne punkty dobrze pokazuje, że mamy do czynienia z niekompatybilnymi ujęciami moralności: funkcjonalistyczną moralnością bez umysłu (świadomości, woli itd.; p. 4.2.1) i moralnością podmiotu w sensie metafizycznym, wyposażonym w rozmaite władze (4.2.2). O ile punkt o moralności funkcjonalnej odwołuje się do współczesnych myślicieli (Floridi, List, Dennett), o tyle punkt o moralnym podmiocie działającym jest głęboko osadzony w historii: od Arystotelesa do K. Wojtyły, A. MacIntyre'a, Ch. Taylora i P. Strawsona. Jak podkreślałam wyżej, historyczne osadzanie problemów uważam za dużą zaletę pracy. Doktorant bardzo dobrze podsumowuje dyskusję, sporządzając listę aspektów przeoczonych lub ignorowanych przez funkcjonalistyczne podejście do moralności (s. 125-127). Zajmuje też wyraźne stanowisko: „zredukowanie moralnej podmiotowości do zachowania czy wskaźników efektywności (*performance indicators*) może być uznane za wysoce niekompletny obraz, który może objąć sztuczne podmioty działające, ale który nie ujmuje prawdziwej natury moralności" (s. 128). Zgadzam się z Autorem, a co więcej, widzę w tym pewną uniwersalną tendencję do zmieniania pojęć tak, by objęły pożądane elementy. Jest to inżynieria pojęciowa, która ostatecznie jest przykrawaniem pojęć na potrzeby pozapoznawczych celów. Doktorant tak tego nie formułuje, ale sądzę, że jego rozważania wspierają moją tezę. Następny punkt poświęcony dużym modelom językowym i zawartych w nich (pozornie) modelach świata stanowi ilustrację powyższej tezy. Autor pokazuje, że nie da się pojęć wartościujących jednoznacznie przełożyć na ciągi zer i jedynek, co jest niezbędne dla napisania algorytmu, a AI nie „zrozumie" ludzkich wartości, ponieważ nie uczestniczy w ludzkim życiu. Ostatecznie problem sztucznych podmiotów moralnych okazuje się problemem metafizycznym (zob. bardzo ważna strona 132). Zgódźmy się z Autorem, że mówienie o moralnych maszynach czy o sztucznych moralnych podmiotach działających jest co najwyżej metaforyczne, a przedsięwzięcie *value alignment* z góry skazane na porażkę, gdyż wiedza moralne jest częściowo wiedzą niejawną i że nie może być raz na zawsze skodyfikowana. Problem wszak pozostaje: jak zagwarantować, że autonomiczne działania AI będą etycznie akceptowalne? Czy jest realna alternatywa dla dyskutowanych w dysertacji podejść? Doktorant zauważa tę kwestię, ale jej systematycznie nie podejmuje, choć zauważa, iż powyższe rozstrzygnięcie nie powinno prowadzić do zarzucenia badań w etyce maszyn.

Ostatni punkt 3. rozdziału nosi tytuł „Kreatywny charakter działania moralnego". Główną tezę da się streścić następująco: moralne działanie jest zarazem autokreacją i współkreowaniem otaczającej rzeczywistości. Zdaniem Autora dysertacja przez to wprowadza w debatę o moralności AI perspektywę nieobecną wcześniej. Autor ma rację jeśli chodzi o anglosaskie debaty, ale nie te toczone w polskiej tradycji. Sam Autor sięga do J. Tischnera i K. Wojtyły, a rozważania warto by uzupełnić o myśl M.A. Krąpca czy T. Stycznia. Są tam doskonałe analizy ludzkiej decyzji jako aktu moralnego, w którym odbywa się współdziałanie rozumu i woli i który ostatecznie prowadzi do sądu: „w tej oto chwili i w tej sytuacji ja powinnam zrobić to a to", a potem sądu: „zrobię to a to". Sądy te nie są dedukcją z jakiś ogólnych zasad, ani indukcją z przypadków, ale właśnie jednostkowym sądem twórczym. Sądem tym ustanawiam siebie jako źródło działania. Ludzka wolność przejawia się zaś w tym, że mogę odrzucić własny sąd „powinnam" i wydać sąd „zrobię całkiem coś innego". Doświadczenie, że wiem, co powinnam zrobić, a robię coś innego znane jest zapewne każdemu.

Taka decyzja ma zawsze konsekwencje nieprzechodnie (buduję siebie jako pewną osobowość) i przechodnie (w świecie uruchamiam nowy łańcuch przyczynowo skutkowy). Aplikację tego rozumienia decyzji do świata AI przedstawiałam w tekstach, m.in. „Morality in the AI World, "Law and Business" (Sciendo) 1(2021), 44-49; „W cieniu rozkwitających AI", w: *Człowiek wobec sztucznej inteligencji,* red. Jerzy Kaczorowski, Poznań 2025, s. 49-58. Krótko mówiąc, nie jest tak, że kwestia kreatywności aktu decyzyjnego (a on ma zawsze wymiar moralny) nie była dyskutowana w kontekście AI. Natomiast zgadzam się z Doktorantem – zresztą podobny był wynik moich analiz – że jeżeli moralne decyzje i idące za nimi działania są twórcze, to idea sztucznego podmiotu moralnego stoi przed poważnym wyzwaniem: pytanie nie dotyczy po prostu tego, czy maszyny mogą wykonać normy moralne czy optymalizować rezultaty etycznie istotne, ale czy są zdolne do autodeterminacji i twórczej odpowiedzi na napotkane wartości (zob. s. 144). Doktorant zajmuje tu jasne stanowisko: sztuczne podmioty działające nie są autentycznie, prawdziwie moralne (s. 146) – dodajmy: o ile nie „przykroi" się pojęcia moralności, ignorując tradycję filozoficzną. Odrzucenie owej postawy „przykrawania" i zanurzenie rozważań w całej filozofii uważam za ważny walor rozprawy. Autor proponuje też alternatywę: zamiast patrzenia na sztuczne systemy jako potencjale moralne podmioty działające, lepiej jest na nie patrzeć jako na „narzędzie moralne" pozwalające człowiekowi lepiej rozumować w sferze moralnej, choćby identyfikować kwestie moralne, przewidzieć konsekwencje, a nawet proponować rozwiązania (s. 146). Są ku temu dobre argumenty (choćby odróżnianie przetwarzania danych od semantycznego rozumienia), choć dla ich formułowania spora część rozważań o historii pojęcia kreatywności nie jest potrzebna. Słusznie Pan Mróz przywołuje metafizyczną zasadę *operari sequitur esse*, ale chyba lepszym tłumaczeniem byłoby chyba „action follows from the being", a nie „action reveals existence" (s. 143), bo łacińskie sformułowania głosi, że każda rzecz działa zgodnie z tym, czym jest (jakim bytem jest).

Zakończenie streszcza główne myśli dysertacji. Doktorant doskonale wypunktował główny problem obszaru AI: oto zauroczeni możliwościami AI, „obniżyliśmy poprzeczkę" rozumienia, czym jest moralna podmiotowość i moralne działanie, i gotowi jesteśmy poszerzyć istotne pojęcia (zadowalając się funkcjonalną świadomością, funkcjonalną moralnością) tak, żeby i AI zaliczyć do podmiotów moralnych. Słusznie Autor odrzuca ten krok, ale szkoda, że nie poświęcił punktu pracy, by w sposób systematyczny pokazać konsekwencje tego kroku, gdyż jest to kwestia istotna kulturowo, a wszak kultura ma być „miejscem teologicznym". W zakończeniu pojawia się kwestia rozumienia odpowiedzialności, ale nie jest to kwestia jedyna – należałby spojrzeć na relacje z AI (mój przyjaciel robot?) czy też uczestnictwa robotów w społeczeństwie. Krótko mówiąc, należałoby przyjrzeć się rozmaitym konsekwencjom uznania AI za podmioty moralne. Autor widzi tę kwestię, – twierdzi bowiem, że sposób w jaki formułujemy debatę o rozwoju technicznym ma istotne znaczenie, ponieważ najpierw my kształtujemy technikę, a potem ona nas – ale jej nie podejmuje (s. 155). Tu wraca kwestia nasycenia techniki wartościami, o której wyżej pisałam. Słusznie natomiast Autor podkreśla, że próby budowania „moralnych maszyn" ostatecznie prowadzą nas do kwestii, które nie mogą zostać rozwiązane w obszarze techniki. Jeśli się z tym zgodzimy – ja z pewnością tak – to staje się to ukrytą krytyką edukacji, z której eliminuje się jako „nieużyteczne" przedmioty humanistyczne, w tym filozofię. Doktorant wraca w „Zakończeniu" do teologii, która „może wnieść wkład" w debatę o AI – ale tego nie pokazuje, bowiem to, do czego sięga w rozważaniach należy do filozofii. Cytat z Papieża Franciszka otwiera natomiast nową kwestię:

wagę rzeczy małych a pięknych, które nie dadzą się zamknąć w algorytmach. A może dadzą? Roboty empatyczne rozkwitają. Natomiast ma rację Pan Mróz twierdząc, że rozwój AI nakazuje zapytać, co znaczy być człowiekiem (s. 155). Szkoda, że na fundamencie dysertacji Pan Mróz nie pokazał kolejnych obszarów badawczych – poza pytaniem o człowieka. Doktorant jest natomiast ostrożny w formułowaniu wniosków i na ogół zaznacza, że taki jest stan obecny AI i dalszy rozwój może to zmienić. Podzielam tę ostrożność, ale sądzę, że uzyskanie przez AI świadomości nie rozwiąże kwestii, bo zrodzi się nowa: czy posiadanie świadomości czyni AI *osobą*, a ostatecznie to o bycie osobą chodzi, a nie po prostu o bycie człowiekiem jako przedstawicielem gatunku.

Doktorant wykazuje dużą świadomość metodologiczną i dobrze panuje nad kolejnymi zagadnieniami, ale przekład tego na formalną konstrukcję pracy jest mocno niezadowalający Dla przykładu punkt 1.1 ma 11 podpunktów, a za to punkty 1.2 oraz 2.7 i 2.9 po jednym podpunkcie – a dobre praktyki edytorskie wskazują, że jednego podpunktu się nie wydziela. Pan Mróz zakłada też sporą wiedzę u czytelnika, bo nie tłumaczy pewnych pojęć jakoś istotnych dla rozważań, np. *biased algorithms* (s. 92), *superintelligence* (s. 92), czy nawet kontrowersji wokół *intelligence* (choć sam problem zauważa (s. 12 i 14), a to okaże się ważne, gdy pojawi się osobliwa definicja inteligencji (s. 94), *social robots*, czy też Anthropic. Oczywiście można sprawdzić w Internecie, ale drobny przypis ułatwiałby lekturę.

Język dysertacji jest dobry, konstrukcja zdań jasna. Dobra jest też konstrukcja rozważań: od naszkicowania kontekstu, przez analizy podmiotowości i podmiotu działającego oraz sztucznych „agentów" (wraz z kwestią etyki maszyn) aż do postawienia pytani, czy moralność może być zalgorytmizowana. Są co prawda fragmenty, które można by pominąć (jak o kreatywności w ogóle), ale nie zakłócają one w sposób zasadniczy toku wywodów. Dużo do życzenia pozostawia redakcja pracy. W osobliwy sposób używane jest wytłuszczenie w tekście (np. s. 76-77), a w innych miejscach w podobnych konstrukcjach nie ma, podobnie czasem w listach główne terminy są wyboldowane (np. s. 89), a czasem nie są (s. 104). Są wyrażenia, które sprawiają wrażenie cytatów (pojawia się cudzysłów), ale brakuje odnośnika, np., s. 115, 121, czy 144. Równie osobliwie używane są odnośniki, np. Doktorant przywołuje Arystotelesa i *Etykę Nikomachejską,* a pojawia się odnośnik do Talbota (s. 118), odwołuje się do Tomasza i jego *Sumy Teologicznej,* a w pewnym momencie odnośnik do Pope'a (s. 118), pisze o Kancie i nawet używa zwrotu „he defines", a odnośnik jest do Davisa i Steibocka (s. 119), pisze o Wojtyle i nawet wygląda jakby był cytat, a odnośnik do Williamsa i Bengsona bez podanej strony. Są to tacy autorzy, że należy sięgnąć do ich własnych dzieł, a nie literatury o nich – a tak interpretuję owe osobliwe odnośniki.

Na strasznym poziomie jest redakcja bibliografii – choć wykaz literatury jest zadowalający pod względem merytorycznym. W wielu pozycjach brak elementarnych danych bibliograficznych, np. Runciman, Przegalińska, Nietzsche, MacIntyre, Levinas, Heidegger, Bostrom czy Boden (nie wypisałam wszystkich przypadków); część danych bibliograficznych jest zupełnie osobliwa, jak Floridi czy Clark: „In Oxford University Press e-Books", choć są standardowe dane dostępne w Internecie; są niekonsekwencje w podawaniu tytułów czasopism: raz kursywą, raz nie; brak stron przy wielu artykułach… niekiedy są nawet skompilowane dwie różne pozycje (zob. np. dane artykułu J. Bryson dostępne w internecie i zamieszczone w bibliografii). Nie wymieniam wszystkich dostrzeżonych przeze mnie błędów, bo już te wskazane pokazują, że Doktorant nie dochował staranności przy sporządzaniu bibliografii. Jeśli

było to robione ze wsparciem AI (co mogłoby sugerować wyjaśnienie kończące wstęp), to doprawdy sztuczna inteligencja nie powinna zastępować naturalnej.

Podsumowując: dysertacja jest rzetelną i informatywną rekonstrukcją jednej z kluczowych debat w obszarze AI: uznaniu AI za sztuczny podmiot moralny (sztucznego agenta moralnego). Z pożytkiem mogli przeczytać tę pracę i filozofowie, i inżynierowie. Rekonstrukcja oparta jest na merytorycznie istotnej literaturze, a za ważną zaletę rozważań uważam ich historyczne osadzenie oraz pokazanie, że rozwiązywanie tego problemu wymaga nie tylko wiedzy technicznej, ale także – a może przede wszystkim – przede wszystkim filozoficznej. Brakuje systematycznego rozważenia konsekwencji, które uznanie istnienia sztucznych agentów moralnych miałoby dla rozumienia człowieka, społeczeństwa czy kultury. Brakuje też owego teologicznego spojrzenia na kwestię moralnych agentów – co zapewne ujawniłoby się wyraźnie przy rozważaniu konsekwencji „uczłowieczania maszyn". Natomiast doskonale ujawnia się kwestia „przykrawania" pojęć moralności czy działania. Pytanie: dlaczego tak się dzieje? Wykracza poza zakres dysertacji, ale kulturoznawca czy politolog powinni je podjąć, podobnie jak filozof mógłby podjąć systematyczne badania nasycenia AI wartościami (np. nasycenia promocyjnego, zob. wskazany wcześniej tekst R. Lizuta). W tym sensie dysertacja jest płodna poznawczo. Doktorant nieczęsto ujawnia swe poglądy, ale ostatecznie formułuje jasne stanowisko co do możliwości uznania AI za moralny podmiot działający (moralnego agenta). Konstrukcja pracy jest metodologicznie zdyscyplinowana, natomiast strona edytorska pozostawia wiele do życzenia, podobnie jak konstrukcja bibliografii.

Biorąc pod uwagę powyższą ocenę rozprawy doktorskiej stwierdzam, że spełnia ona wymogi postawione w *Ustawie o stopniach naukowych i tytule naukowym* i wnioskuję o dopuszczenie mgra Macieja Mroza do dalszych etapów postępowania doktorskiego.

Ewa Agnieszka Lekka-Kowalik
Lublin, 19 stycznia 2026