

Dr hab. Maciej Bała, prof. ucz.

UKSW

Maciej Mróz, „Moralność sztucznych agentów”, rozprawa doktorska, UAM, Wydział Teologiczny, Poznań 2025

Rozprawa doktorska Macieja Mroza podejmuje pytanie, które w filozofii techniki i etyce AI ma dziś rangę pytania granicznego: czy i w jakim sensie sztuczne systemy decyzyjne mogą być podmiotami moralnymi, a zatem bytami, którym przypisuje się nie tylko skuteczność działania, lecz także odpowiedzialność, normatywną sprawczość oraz zdolność do rozumienia powinności. Praca jest formalnie osadzona w dyscyplinie nauk teologicznych, lecz jej rdzeń argumentacyjny stanowi filozoficzna analiza pojęć: agencji, intencjonalności, wolności, odpowiedzialności i normatywności. Autor świadomie przyjmuje metodę egzegetyczno-krytyczną, rekonstruuje stanowiska, a następnie je oceniając za pomocą narzędzi takich jak rekonstrukcja logiczna, kontrprzykłady i równowaga refleksyjna, uzupełniając ten aparat elementami podejścia fenomenologicznego i hermeneutycznego.

Tym, co nadaje pracy szczególną wartość filozoficzną, jest konsekwentna próba odsłonięcia ukrytych założeń, które stoją za językiem „moralnych maszyn” i „agentów moralnych”. Autor pokazuje, że spór o moralność sztucznych agentów nie jest w pierwszym rzędzie sporem technicznym, lecz sporem o to, czym jest moralność oraz jaki typ bytu może być jej nosicielem. Jest to w istocie spór metaetyczny i antropologiczny, w którym powracają klasyczne napięcia między zewnętrżnością i wewnętrżnością działania, między opisem zachowania a rozumieniem sensu, a także między regułą a żywą praktyką. Już sama diagnoza, że w obszarze AI terminy „wartość” czy „dobro” zmieniają znaczenie w zależności od rejestru (etycznego lub inżynierskiego), prowadzi autora do podstawowego pytania: co to znaczy, że algorytm jest „dobry”, i czy „dobro” może zostać zredukowane do parametru.

W rozdziale pierwszym autor buduje fundament pojęciowy poprzez analizę definicji sztucznej inteligencji oraz konsekwencji, jakie niosą. Ta część, choć ma charakter wprowadzający, spełnia funkcję filozoficznie istotną: ujawnia, że definicje AI kształtują wyobrażenia o tym, co jest (a co nie jest) inteligencją i sprawczością. Rozważania o teście Turinga i o współczesnych modelach generatywnych nie służą tu jedynie historycznej narracji, lecz stają się punktem wyjścia do krytyki behawioryzmu poznawczego, który skłonny jest utożsamiać inteligencję i moralność z ich zewnętrzną manifestacją. Autor słusznie wydobywa ryzyko epistemiczne polegające na tym, że płynność językowa i przekonujące uzasadnienia

generowane przez systemy mogą wywoływać wrażenie rozumienia i intencjonalności. W tym miejscu praca włącza się w klasyczną debatę o rozróżnieniu między kompetencją a zrozumieniem, przypominając, że pomyłka w tym rozróżnieniu nie jest niewinna, lecz może stać się „coraz bardziej niebezpieczna”, ponieważ przenosi ciężar z rozumienia na perswazję i skuteczność.

Rozdział drugi rozwija pojęcie sztucznej agencji, przy czym jego znaczenie filozoficzne polega na wykazaniu, że „agencja” nie jest terminem jednowymiarowym. Autor opisuje stopniowalność i progowość sprawczości oraz rozróżnia agencję funkcjonalną, zbiorową i rozproszoną, a w tle stale powraca napięcie między sprawczością jako zdolnością inicjowania skutków a sprawczością jako zdolnością bycia odpowiedzialnym za sens i normatywny ciężar działania. To rozróżnienie ma kluczowe znaczenie dla całej rozprawy, ponieważ pozwala rozbroić popularny skrót myślowy: skoro system „działa autonomicznie”, to „jest odpowiedzialny”. Autor pokazuje, że autonomia techniczna i autonomia moralna nie są tym samym, a próby ich utożsamienia są konsekwencją antropomorfizacji oraz funkcjonalistycznego przesunięcia kryteriów.

Najbardziej wyraźne starcie filozoficzne rozprawa rozgrywa w rozdziale czwartym, gdzie autor przeciwstawia funkcjonalistyczną koncepcję moralności „standardowym” ujęciom sprawczości moralnej. Funkcjonalizm, jak pokazuje autor, zakłada, że do agencji moralnej wystarczą odpowiednie zdolności funkcjonalne lub obserwowalne zachowania, nawet jeśli brak świadomej intencji, wolnej woli w sensie mocnym czy cech duchowych. Wersja inżynierska funkcjonalizmu jest atrakcyjna, ponieważ pozwala „ominać” spory metafizyczne i skupić się na tym, co mierzalne. Autor bardzo trafnie zauważa jednak, że ta oszczędność ontologiczna ma cenę: usuwa z pola widzenia to, co w moralności wydaje się „nadające znaczenie moralne” samej sprawczości, czyli prawdziwe rozumienie, autentyczny wybór, relację do wartości, emocjonalno-egzystencjalne zaangażowanie oraz odpowiedzialność w sensie nie tylko przypisania skutków, lecz poczucia powinności.

W tej linii krytyki pojawia się szczególnie mocna intuicja, którą autor wiąże z pojęciem „moralnych zombie”: bytów, które wykazują zachowania moralne bez „życia moralnego”, a więc bez wewnętrznej perspektywy normatywnej. Ta figura jest filozoficznie nośna, ponieważ pokazuje, że sam opis „poprawnego” zachowania nie przesądza o tym, czy mamy do czynienia z moralnością, czy tylko z jej imitacją. Z perspektywy metaetycznej jest to powrót do pytania, czy normatywność da się zredukować do regularności behawioralnej albo do optymalizacji, czy też ma ona nieredukowalny wymiar „powinności” i „uzasadnienia”, którego nie da się uchwycić w samym wejściu-wyjściu systemu.

Na tym tle bardzo dobrze wypada analiza moralnego testu Turinga. Autor nie traktuje tej koncepcji jako ciekawostki historycznej, lecz jako modelowego przykładu tego, jak behawioryzm próbuje obejść spory etyczne poprzez przeniesienie kryterium moralności na nieodróżnialność konwersacyjną w rozmowie o moralności. Autor rekonstruuje założenia MTT i pokazuje, że problem testu nie leży wyłącznie w jego praktycznych ograniczeniach, lecz w samym pomysle zastąpienia moralności jej perswazyjnym opisem. Wykorzystana krytyka Arnolda i Scheutza wzmacnia ten punkt: test oparty na naśladownictwie miesza zdolność mówienia o moralności z posiadaniem kompetencji moralnej, premiuje czarną skrzynkę i z definicji nie wymaga integralności procesu od percepcji przez ocenę po działanie; co więcej, tworzy paradoksalną zachętę do „oszustwa” jako warunku zdania testu, a więc czyni z moralności praktykę, której miarą jest zdolność do zwodzenia rozmówcy. W rezultacie autor trafnie wykazuje, że kryteria behawioralne nie tylko nie wystarczają, ale mogą wręcz deformować rozumienie moralności.

Istotnym pogłębieniem filozoficznym jest włączenie do dyskusji perspektywy Lévinasowskiej. Autor wydobywa, że etyka u Lévinas’a nie jest przede wszystkim funkcją zdolności podmiotu do rozumowania i wyboru, lecz strukturą odpowiedzi na wezwanie Innego, która w pewnym sensie „podważa moją wolność” i ma charakter asymetryczny. W tym świetle moralna podmiotowość jawi się raczej jako wrażliwość i podatność na roszczenie Innego, a nie jako zestaw sprawności obliczeniowych. Włączenie tej perspektywy jest cenne, ponieważ demaskuje ograniczenie wielu projektów etyki maszyn: nawet jeśli maszyna potrafi obliczyć „właściwy” wybór, pozostaje pytanie, czy w ogóle znajduje się w egzystencjalnej sytuacji etycznej, w której rodzi się sumienie i odpowiedzialność. Autor sugeruje, że symulacja troski nie jest jeszcze byciem „roszczonym” przez Innego i że to właśnie ten wymiar relacyjno-fenomenologiczny ginie w funkcjonalistycznych listach kontrolnych.

Jednocześnie rozprawa nie poprzestaje na krytyce. Autor uczciwie rekonstruuje również argumenty bardziej „przychylne” maszynom, w tym inspiracje stanowiskiem Dennetta, gdzie moralność rozumiana jest jako praktyczne rozwiązanie problemów koordynacyjnych, a wolna wola w sensie kompatybilistycznym ma postać zdolności do samokontroli i refleksji, nie zaś metafizycznej „niezdeteminowanej” mocy wyboru. Z filozoficznego punktu widzenia jest to ważne, ponieważ pozwala uchwycić, że spór o moralność maszyn zależy od tego, jak rozumiemy moralność: czy jako system reguł i kompetencji, czy jako egzystencjalną formę życia, w której powinność jest wewnętrznie przeżywana.

Najbardziej oryginalny wkład autora ujawnia się jednak w części poświęconej twórczemu wymiarowi działania moralnego. Autor przekonująco pokazuje, że dominujące w

debacie AMA – deontologia, konsekwencjalizm i etyka cnót – bywają w praktyce implementacyjnej upraszczane do schematów opartych na regułach i na optymalizacji działań, podczas gdy znaczna część tradycji filozoficznej rozumie moralność nie jako mechaniczne stosowanie zasad, lecz jako twórczą odpowiedź na sytuacje bezprecedensowe. Włączenie Bergsona, Tischnera i Wojtyły pozwala autorowi opisać akt moralny jako proces samostworzenia, w którym podmiot nie tylko realizuje normy, lecz także kształtuje własną tożsamość moralną. Autor trafnie wydobywa u Bergsona rozróżnienie między moralnością „zamkniętą” a „otwartą” oraz związek autentycznego działania moralnego z kreatywną emocją i generowaniem nowych możliwości, a nie z presją społeczną czy kalkulacją. Wątek Tischnera wzmacnia rozumienie moralności jako „samokompozycji”, gdzie wybory etyczne tworzą człowieka, a nie tylko skutki, zaś analiza Wojtyły pozwala uchwycić moralność jako szczególnie typ przyczynowości związanej z samostanowieniem, która przekracza determinizm materialny i prostą optymalizację.

Ta część rozprawy ma dalekosiężne implikacje dla etyki AI. Jeżeli bowiem moralność w istotnym sensie zawiera moment twórczego przekraczania danych warunków, to projekt „obliczania moralności” staje się projektem kategorialnie błędnym: miesza stosowanie reguł z wytwarzaniem sensu i wartości. Autor konsekwentnie sugeruje, że maszyny mogą reprodukcować wzorce i generować pozory nowości, ale prawdziwa kreatywność – rozumiana jako przekroczenie algorytmicznej rekombinacji – może okazać się dla nich nieosiągalna, a przez to także pełnia moralnego działania pozostaje poza ich zasięgiem. To rozpoznanie stanowi mocny, filozoficznie ambitny argument przeciwko redukcjonizmowi, który próbuje sprowadzić moralność do zestawu procedur.

W tym miejscu warto podkreślić jeszcze jedną zaletę pracy. Autor pokazuje, że nawet jeśli udałoby się zbudować systemy zachowujące się „moralnie”, pozostaje problem normatywnej motywacji i niezawodności, co w debatach o alignment i „pośredniej normatywności” ujawnia się jako trudność zapewnienia, by system realizował zamierzone cele bez „niezablokowanego sposobu wykorzystania” czy błędnej interpretacji procesu. Filozoficznie oznacza to, że moralność nie jest jedynie kwestią celu, lecz także sposobu bycia w świecie, a więc formy życia, w której sens powinności nie jest zewnętrznie narzuconą funkcją.

Choć rozprawa Macieja Mroza prezentuje wysoki poziom erudycji filozoficznej i imponującą znajomość współczesnej literatury z zakresu etyki sztucznej inteligencji, niektóre jej elementy mogą stać się przedmiotem dalszej dyskusji krytycznej. W pierwszej kolejności należy zwrócić uwagę na znaczną rozbudowę części definicyjno-kontekstualnej, zwłaszcza w

rozdziale pierwszym. Obszerność analiz dotyczących definicji sztucznej inteligencji, typologii systemów oraz debat regulacyjnych, choć merytorycznie poprawna i rzetelna, momentami przesłania zasadniczy problem filozoficzny pracy, jakim jest status moralny sztucznych agentów. Czytelnik może odnieść wrażenie, że autor zbyt długo pozostaje na poziomie porządkowania stanowisk, zanim przejdzie do właściwej argumentacji normatywnej, co osłabia dynamikę wywodu.

Pewną ambiwalencję budzi również sposób, w jaki autor korzysta z perspektywy fenomenologicznej i personalistycznej. Odwołania do Lévinas'a, Tischnera czy Wojtyły są filozoficznie inspirujące, jednak w niektórych miejscach pozostają raczej na poziomie ilustracji niż pełnoprawnego sporu argumentacyjnego z etyką maszyn. Czytelnik może odczuć niedosyt bardziej systematycznego pokazania, w jaki sposób kategorie takie jak twarz Innego, samostanowienie czy twórczy akt moralny miałyby stanowić nie tylko krytykę, lecz także precyzyjne kryteria negatywne, pozwalające jednoznacznie wykluczyć możliwość sztucznej podmiotowości moralnej.

Autor konsekwentnie broni stanowiska, zgodnie z którym akt moralny posiada nieredukowalny wymiar twórczy i sytuacyjny, jednak teza ta – przyjęta w mocnej postaci – prowadzi do pytania, czy analogiczne zastrzeżenia nie należałoby zgłosić również wobec części ludzkich praktyk moralnych, które w rzeczywistości opierają się na schematach, nawykach i regułach. Innymi słowy, rozprawa mogłaby zyskać na precyzyjniejszym rozróżnieniu między tym, co w moralności konstytutywnie twórcze, a tym, co w praktyce bywa rutynowe, a mimo to pozostaje moralnie relewantne.

Wreszcie, można odnotować pewne napięcie między krytyką antropomorfizacji a wykorzystaniem relacyjnych ujęć etyki, które same w sobie opierają się na fenomenologii spotkania i doświadczenia sensu. Autor słusznie przestrzega przed przypisywaniem maszynom cech osobowych na podstawie powierzchownych sygnałów, jednak nie zawsze w pełni wyjaśnia, w jaki sposób odróżnić „pozór relacji” od relacji moralnie znaczącej, zwłaszcza w kontekstach społecznych, w których ludzie realnie reagują na systemy sztuczne w sposób emocjonalny i normatywny. To napięcie nie podważa głównej tezy pracy, ale wskazuje na obszar, który mógłby zostać rozwinięty w dalszych badaniach.

Zgłoszone uwagi krytyczne nie podważają zasadniczej wartości rozprawy, lecz świadczą o jej ambicji teoretycznej i otwartości na dalszą debatę filozoficzną. Praca nie zamyka problemu moralności sztucznych agentów, lecz stawia go na poziomie, który wymaga dalszej, pogłębionej refleksji nad samą naturą moralności, sprawczości i odpowiedzialności.

Podsumowując, rozprawa Macieja Mroza jest pracą dojrzałą i filozoficznie ambitną. Jej największą siłą jest pokazanie, że pytanie o moralność sztucznych agentów nie rozstrzyga się na poziomie „czy system potrafi dać dobrą odpowiedź”, lecz na poziomie tego, czym jest moralność jako taka: czy jest ona obliczalnym zbiorem reguł, czy też formą życia, w której normatywność jest wewnętrznie przeżywana, relacyjna i twórcza. Włączenie perspektyw Bergsona, Tischnera, Wojtyły oraz Lévinas’a pozwala autorowi zbudować krytykę redukcjonizmu, która nie jest jedynie negacją programu etyki maszyn, lecz pozytywnym doprecyzowaniem pojęcia moralności. Z tego względu rozprawa spełnia wymagania stawiane pracy doktorskiej, wnosi oryginalny wkład w debatę o etyce AI.

Ja, niżej podpisany stwierdzam, że recenzowana rozprawa doktorska Maciej Mroza spełnia warunki określone w art. 13.1 Ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Dz.U. nr 65 poz. 595 z późn. zmianami) i wnioskuję do Rady Dyscypliny Nauki Teologiczne o dopuszczenie Macieja Mroza do dalszych etapów przewodu doktorskiego.