Ks. prof. dr hab. Piotr Roszak
Wydział Teologiczny UMK

Toruń, dn. 30.03.2022

Recenzja rozprawy doktorskiej
mgra Łukasza Sarowskiego
pt. *Robot humanoidalny jako podmiot życia społecznego. Problemy filozoficzne i społeczne,* Lublin 2022.

Jak zauważają niektórzy filozofowie nauki, w przypadku realizacji przedsięwzięć naukowych można wyróżnić dwa rodzaje badaczy. Jedni przypominają „mechaników", którzy swoją rolę ograniczają do prostego naprawienia czy poprawienia teorii/prawa, gdy zostanie zaobserwowane, że coś nie działa. Ale istnieje inna grupa ludzi, którzy przypominają „pionierów", odkrywających nowe przestrzenie i zastosowania. Mierzą się z innymi wyzwaniami i trudno od nich oczekiwać wyczerpujących opisów, są pod wrażeniem rzeczywistości i skupiają się na tym, co kluczowe dla wypracowania nowych rozwiązań. W świetle tego rozróżnienia, wydaje mi się, że rozprawa doktorska p. mgra Łukasza Sarowskiego wpisuje się w ten drugi rodzaj pracy naukowej. Prezentuje bowiem nowe obszary refleksji filozoficznej związane z obecnością robotów w społeczeństwie, o coraz większej inteligencji i autonomiczności, zdolnych nie tyle do wiernego podążania za algorytmami, co do pewnej kreatywności i podejmowania decyzji na podstawie wielu zmiennych czynników. Nie chodzi jednak o same nowe możliwości maszyn, ale włączenie się w jedną z najnowszych debat dotyczących podmiotowości robotów, sposobów, w jakie wchodzą w interakcje i związanej z tym refleksji etycznej oraz antropologicznej (kim jest człowiek można poznać przez odniesienie go do tego, co nie jest człowiekiem). To zaś pociąga za sobą łączenie często odległych światów: Doktorant musi rozumieć wyzwania przemysłu 4.0 i związanej z nim automatyzacji, ale także wykazać się dobrą znajomością klasycznych debat filozoficznych.

Aby podjąć się tego ambitnego zadania potrzeba wypracowania swoistej „mapy pojęciowej", bo choć zawsze odkrywcy nowych ziemi nazywali „starymi terminami" to, co odkrywali jako nowe (wystarczy wrócić do historii i zachowań np. hiszpańskich

konkwistadorów z XVI w.), to jednak dla rozwoju dyskusji i znakiem jej pogłębienia jest precyzacja stosowanego aparatu pojęciowego. Temu wyzwaniu poświęcona jest praca doktorska, która przez to nie przypomina klasycznych dysertacji, do jakich jesteśmy przyzwyczajeni w filozofii, a w których w centrum stawia się konkretną hipotezę badawczą, będącą propozycją interpretacyjną danego zjawiska czy sposobem rozwiązania jakiegoś problemu. W przypadku rozprawy Ł. Sarowskiego tego nie znajdziemy (choć Doktorant formalnie stawia hipotezę, niemniej jest ona zbyt ogólna – s. 11: „stawiam hipotezę badawczą, że zakres ich [własności podmiotowości społecznej – PR] stosowalności do robotów humanoidalnych decyduje o przyznaniu im statusu podmiotu życia społecznego"), nie proponuje jednej nowej teorii, ale chce dokonać oceny prowadzonych dyskusji filozoficznych w ostatnim czasie, aby przekonać się na jakim etapie jest już uzgodniona owa „mapa" pojęć. Nie jest to jednak „wypracowanie", co „zweryfikowanie" już istniejącej sieci pojęć. To z pewnością tłumaczy mniejszą niż zazwyczaj objętość pracy, a także jej charakter, w której nie tyle Doktorant stawia swoje hipotezy, co krytycznie ocenia istniejącą literaturę przedmiotu. Zalety i wyzwania tego podejścia warto szczegółowo omówić.

1. Ocena formalna

Dysertacja doktorska Ł. Sarowskiego została podzielona na pięć rozdziałów, przy czym ich powiązanie tematyczne nie jest ciągiem przyczynowo-skutkowym, ale ma charakter promienisty: to kolejne tematy jakie rodzi refleksja nad podmiotowością robotów, które Doktorant omawia szczegółowo. W ten sposób musi łączyć wątki wydawałoby się nie zawsze wprost związane z prowadzonym badaniem, bo nie wszyscy dostrzegają związek bezpośredni między podmiotowością a wyzwaniami kryjącymi się za problemami testów na inteligencję sztucznych systemów poznawczych, które autor wnikliwie objaśnia.

Dokonywane przez Doktoranta refleksje zdradzają dobrą znajomość filozofii, zwłaszcza współczesnej, na której opiera większość swoich rozważań, traktując bardzo – może nawet zbyt ogólnie – filozofię antyczną i średniowieczną. Proponowane przez Doktoranta podsumowania stanowisk filozofów średniowiecznych nie są dokładne i bazują na pewnych topikach, zamiast na pogłębionej analizie. Widać to wyraźnie, np. w

rozważaniach nad antropologią Arystotelesa (s. 14), gdzie kwestie rozumu czynnego i biernego, podsumowuje stwierdzeniem, że są konsekwencją przyjęcia „ogólnego rozumienia bytu, który istnieje". W ogóle, wydaje się dość niebezpieczne posługiwanie się ogólnie słowem filozofia, co widać w zapowiedziach tematu rozprawy, ale i rozważaniach szczegółowych, by przywołać przykład: „W filozofii dyskusję nad autonomicznością sprowadza się do problemu wolności decyzji, woli i wyboru" (s. 85), ale jakiej filozofii? Nie w każdej przecież. Zdecydowanie jednak więcej staranności dokłada Doktorant w przypadku współczesnej filozofii i jej autorów. Ten brak jasnego podkreślenia na bazie jakiej filozofii chce dokonać „weryfikacji" lub „wypracowania" narzędzi pojęciowych dla współczesnej filozofii robotyki jest jedną z trudności dla oceny realizacji zamierzonego celu badawczego. Wiele prób np. podejmuje się ostatnio z hylemorfizmem, czy ogólnie na tle neo-arystotelizmu: Doktoranta jednak interesują inne nurty współczesnej filozofii, a może nawet podejście b. eklektyczne, co czyni z pracy bardziej ambitne przedsięwzięcie, ale jednocześnie wymagające lepszego zakreślenia pola badawczego.

Atutem dysertacji jest dobre oparcie na teorii i praktyce, co widać w harmonijnym łączeniu rozważań filozoficznych z osiągnięciami technologicznymi współczesnej robotyki. Dlatego odnosi się Doktorant do np. modelu karalucha SimRoach2, robota Khepera, wspomina także robota PARO, który miał zmniejszać poziom stresu u pacjentów, a także co ciekawe dla świata akademickiego przytacza IBM Miss Debater, który jednocześnie dotyczy kwestii uczenia się maszyn, a nie podążania po ścisłych trajektoriach algorytmów, a co prowadzi Doktoranta do rozważań nt. języka naturalnego i w jaki sposób roboty mogą odpowiednio posługiwać się tym językiem.

Doktorant niestety podaje pewne sformułowania, ale bez pokrycia w źródłach, np. str. 85 wymienia różne rodzaje autonomiczności w naukach szczegółowych (autonomia adaptacyjna, behawioralna, przekonań, mentalną, autonomię motywacyjną, autonomię społeczną, autonomię jednostki, autonomię agentów), ale żadnego przypisu, który by potwierdził to przekonanie lub egzemplifikował poszczególne rodzaje . Uwaga Doktoranta, że „[w] perspektywie tytułowego problemu dysertacji ich szczegółowe przedstawienie nie jest konieczne. Chcę jedynie zasygnalizować złożoność zagadnienia, które aplikuje się obecnie w obszar dyskusji nad autonomicznością robotów" rodzi to

pytanie czy rzeczywiście jednak nie warto było tego rozwinąć. Tym bardziej, że w moim przekonaniu Doktorant pokazuje jedynie „gdzie" się pojawia problem autonomiczności, a nie jak go odnosi do kwestii robotów. Tej refleksji niestety brakuje w całej pracy.

Zdarzają się Doktorantowi drobne potknięcia językowe, np. „komputerowa metafora umysłu" (s. 102) – co może być rozumiane wieloznacznie; pojawiają się zwroty ogólne typu: „badacze podkreślają…" (s.111), ale nie wspomina kto konkretnie, zwłaszcza, że Doktorant przytacza potem ciekawy podział komponentów samoświadomości. Można było również lepiej wyjaśnić kwestie etymologiczne, jak w przypadku terminu inteligencja, o czym Autor informuje w następujących słowach: „Widać to już w jej źródłowym znaczeniu wywodzącym się od łacińskiego terminu *intelligentia*, a więc zdolności rozumienia w ogóle, bystrości, pojętności, zdolności rozumienia sytuacji i odpowiedniego na nie reagowania". Zabrakło odwołania do *intus-legere*, czytania pomiędzy, pośród danych i uchwycenia istoty; ponadto, przypis odsyła nie do słownika łacińskiego czy studium z tego zakresu, ale ogólnej literatury.

2. Ocena metodologiczna

Rozprawa jest napisana, zgodnie z deklaracjami na początku dysertacji, jako próba podsumowania prowadzonych już debat, co domaga się przyjęcia pewnej strategii metodologicznej. To wyjaśnia większe oparcie prowadzonych rozważań na istniejącej literaturze, uporządkowanie padających propozycji oraz ich porównanie między sobą: ten ostatni aspekt nie został przez Doktoranta zbytnio rozbudowany, ale obecny z całą pewnością jest w dysertacji.

Podejmowane refleksje nad robotami humanoidalnymi w kontekście ich „podmiotowości" oparte są o analizę poszczególnych własności związanych z podmiotowością społeczną, co jest b. słusznym rozwiązaniem, gdyż rodzaj podmiotowości jaki znamy to przede wszystkim ludzka aktywność. A zatem *per analogiam* stara się Autor określić zakres obowiązywalności tradycyjnych pojęć, które są aplikowane do robotów we współczesnym dyskursie naukowym.

Nie ulega wątpliwości, że Doktorant bardzo dobrze zna aktualną literaturę i porusza się w niej swobodnie: przytacza trzy prawa robotyki Asimova, koncepcje bezpiecznej sztucznej inteligencji Yudkowskiego, ale trudno doszukać się często

własnego zdania Doktoranta na temat głównych sporów. To widać już w formie językowej, gdy mówi np. „Kluczowe znaczenie w tym wypadku może mieć…" – może czy realnie ma? Po tych słowach wskazuje na problematyczne kwestie (s. 99), ale już nie na ich rozwiązanie. A z pewnością to dodałoby wartości propozycjom. Niemniej z metodologicznego punktu widzenia zajmuje obiektywny dystans i widzi siebie w roli filozoficznego „reportera".

W swoich analizach posługuje się niekiedy uproszczeniami, zwłaszcza w koncepcjach filozoficznych – czasem to dobrze, ale przy analizach szczegółowych źle. Przykładem może być sytuacja, gdy omawiając kwestie dialogu (s. 65), przytacza trzy koncepcje z całej historii filozofii (tradycję sokratejsko-platońską; myśl Levinasa i Bubera oraz Gadamera). Wybór nie jest jednak uzasadniony i jak wiadomo istnieje bardzo wiele interesujących koncepcji zarówno średniowiecznych jak i współczesnych. Niekiedy pojawiają się dość naiwne stwierdzenia, jak ze s. 102: „w roku 1956 odbyła się nie mniej ważna konferencja w Massachusetts Institute of Technology, której efektem było powołanie do istnienia nowej dziedziny nazywanej psychologią poznawczą". Wpływ tej konferencji na badania dziś określone jako psychologia poznawcza jest znaczący, ale trudno w świecie naukowym twierdzić, że jakieś sympozjum czy kongres „powołuje do istnienia" dziedziny nauk.

Inne „klucze", jakie stara się przyłożyć do badanego tematu wydają się bardzo obiecujące, jak gdy właściwie przez całą rozprawę Doktorat nawiązuje do podziału G. Simmela na „Ja indywidualne", „Ja uogólnione" i „Ja ogólne" (s. 66): są one podstawą do zarysowania różnic między relacjami osobowymi i nieosobowymi. W tym kontekście warto było przedstawić również całą nową dziedzinę filozofii (i socjologii), jaką rozwija P. Donati (i jego „socjologia relacyjna"), włoski socjolog, gdyż ma ona potencjał eksplikatywny o szerokim zasięgu.

Niemniej jednak trzeba pochwalić systematyczność refleksji Doktoranta, konsekwencje w prowadzeniu analiz, a także troskę, aby trzymać się zasadniczego tematu, nie ulegając pokusie anegdotyczności (co często zdarza się przy okazji prac na temat robotyki, gdyż *case-study* zajmują niekiedy zbyt dużo miejsca w pracach z tego zakresu).

3. Ocena merytoryczna

Z pewnością wartość merytoryczna pracy to nie wypracowanie oryginalnej, nowej koncepcji interpretacyjnej fenomenów, ale określenie kryteriów wyjściowych dla refleksji nad podmiotowością robotów. Dotychczasowe debaty, toczone na różne sposoby – od robo-entuzjastów po robo-sceptyków – cechują się przede wszystkim chaosem semantycznym, a przez to jak się wydaje stoją w martwym punkcie. W zamiarach Doktoranta leży więc ożywienie debaty przez wskazanie na możliwości i ograniczenia pewnych dyskursów filozoficznych, zbadanie ich propozycji i wyważoną ocenę. Pomimo tak ambitnego zadania, Doktorant zostawia niekiedy czytelnika w rozterce: nie wiadomo bowiem co jest wkładem Doktoranta do tej debaty, a co „zebraniem rozproszonych wątków" (co zresztą wprost uznaje za cel pracy w Zakończeniu).

Jak zauważa Ł. Sarowski, tytułowe upodmiotowienie robotów wiąże się z relacyjnością, którą rozważa nie tyle w świetle klasycznych koncepcji, ile współczesnych (np. „teorii aktora sieci", s.29). Słusznie diagnozuje, że „robotyka społeczna" najczęściej opiera się na odnoszeniu zachowań maszyn do cech ludzkich, a więc na swoistym mimetyzmie. Kluczem jednak (i wyzwaniem) pozostaje stwierdzenie w jakim sensie tradycyjne pojęcia (np. sprawczości, podmiotowości, wolności) mogą odnosić się do robotów. Doktorant, aby odpowiedzieć na to pytanie, proponuje zatrzymanie się najpierw nad samym pojęciem podmiotowości. Może być ona rozumiana jako realizowanie funkcji poznawczych albo relacji / doświadczeń, a te może robot gromadzić, co Autor udowadnia odwołaniem choćby do uczenia maszynowego. Ale nie dokonuje tego w taki sam sposób co człowiek i dlatego Doktorant szuka nowych pól do rozważań, aby uchwycić podmiotowość robotów np. przez tzw. „Ja uogólnione" rozumiane jako funkcje i role społeczne pełnione przez podmiot.

Robot, zdaniem Doktoranta, nie musi cechować się autonomicznością w silnym sensie, jaki przypisujemy człowiekowi, ale wystarcza „interakcja" (s. 40); w ogóle zamiast mówić o wolności robotów, warto mówić o autonomii. Zresztą interaktywność i możliwe zasięgi rozumienia tego terminu (który przecież nie ogranicza się do werbalnej wymiany) są sposobem na rozwijanie kolejnych wątków i w ten sposób

zarysowania przed czytelnikiem drogi rozwoju badań nad robotami, które dziś obejmują już emocjonalne roboty jak Kismet.

Trudno się jednak zgodzić z Doktorantem, który twierdzi, że „w średniowieczu pojęcie autonomii nie występuje. Pojawia się ono dopiero w XVI wieku, w kontekście wolności religijnej. Z kolei w XVII i XVIII w. pojęcie to aplikuje się w obszar prawoznawstwa, a następnie uzyskuje ono ogólniejsze znaczenie w antropologii filozoficznej I. Kanta, dla którego oznacza prawo do instytucjonalnego samostanowienia oraz możliwości samo-konstytuowania się człowieka będącego istotą rozumną" (s. 85). Wydaje się to zbyt redukcjonistycznym ujęciem, ponieważ pojęcie autonomii istnieje w średniowieczu, choć nie musi być tak samo definiowane jak w nowożytności. Dla tamtych myślicieli, autonomia jest rozumiana jako zdolność do samo-determinowania podmiotu (bycie *sui iuris*). Być może tego typu sformułowania pojawiają się dlatego, że Doktorant nie sięga w takich przypadkach do szczegółowych rozpraw, ale ogólnych opracowań. Widać to np., gdy rozważa autonomię w naukach społecznych (jej rozumienie), co wyjaśnia posiłkując się *Encyklopedią Pedagogiki*. W podobny sposób Doktorant zachowuje się omawiając skale autonomiczności: autor podaje kilka przykładów, ale właściwie ich nie omawia, porównuje, nie dokonuje oceny (s. 92).

Zaletą dysertacji doktorskiej jest zwracanie uwagi na kwestie, które stanowią dziś podstawowy nurt dyskusji na temat robotów humanoidalnych, takich jak ucieleśnienie poznania, wyznaczający ważny etat rozwoju badań nad robotami, czy kwestia płciowości robotów, która może mieć wpływ na sposób, w jaki wchodzi się z nimi w relacje (badania Tatsuya Nomury). Autor jest dobrze zorientowany w bieżącej dyskusji dot. tzw. płci mechanicznej (Rogera A. Søraa) i potrafi przekonująco argumentować.

Wszystkie rozważania Doktoranta mieszczą się w ramach szeroko rozumianej „roboetyki" (s. 93), przy czym szkoda, że nie wspomina kto jest autorem tego pojęcia, lecz z góry zakłada jego znajomość, a trzeba przyznać, że nie jest to ustalone: etyka maszyn jest bardziej znanym dziś pojęciem. Jeśli ktoś ma ambicję ustalenia pola znaczeniowego, aby zaoferować 'mapę', musi te niuanse rozważać. W jakiej relacji pozostaje ta „roboetyka" to etyki stosowanej? Poza tym, następne zdanie dotyczy już etyki „antropocentrycznej", jak stwierdza autor. Co to za etyka (zwłaszcza jej

„centryczność" na czym polega)? Autor co prawda przytacza opinię Loh, że roboetyka jest subkategorią etyki maszyn reprezentującą nową etykę stosowaną, ale nie wiadomo czy się z tym zgadza. Widać, że czasami odchodzi od problemów filozoficznych na rzecz praktycznych (co samo w sobie nie jest zarzutem) – gdy np. przytacza kodeks EURON i 5 zaleceń dotyczących roboetyki.

Niekiedy w rozprawie pojawiają się mocne sformułowania (filozoficznie), jak to, że samoświadomość jest główną cechą podmiotu (s. 87), ale wówczas co powiedzieć o osobach niepełnosprawnych (lub innymi dolegliwościami klinicznymi)? Czy one są mniej podmiotami? Jak wiadomo toczy się dziś szeroki spór związany z podmiotowością osób niepełnosprawnych (w śpiączce etc.) i ich obecności w życiu społecznym. Zabrakło w pracy rozwinięcia tego aspektu. Z drugiej strony, zastanawiające są refleksje Doktoranta na temat relacji zwierząt do robotów, które jak na prace filozoficzna grzeszą jednak brakiem pogłębionej lektury z historii filozofii: uproszczenia w stosunku do Kanta są właściwe dla ujęć publicystycznych, a nie dysertacji naukowej.

Oceniana dysertacja zostawia niekiedy czytelnika z obietnicami, których jednak nie uwiarygadnia w żaden sposób. Przykładem mogą być słowa, że „[i]nteresująco przedstawiają się jednak dalsze prace w zakresie weryfikacji paradygmatu ucieleśnionego umysłu, zwłaszcza prób aplikacji założeń Lakoffa i Johnsona dotyczących relacji ucieleśnienia względem rozwoju zdolności językowych." – brakuje wyjaśnienia na czym polega ów potencjał. Wyzwaniem jest także kwestia tego, czy w przypadku robotów mamy do czynienia z sytuacji moralnymi? To jest kluczowe pytanie, które stawia sobie autor na str. 96.

Trzeba jednak mocno podkreślić niezaprzeczalne atuty rozprawy, do których warto zaliczyć odniesienia do bieżących debat, jak np. kwestii tzw. *Symbol grounding problem*. Wydaje się, że zdolność do wyrażania (i odczytywania) symboli jest jedną z wyróżniających ludzkich cech, stąd nie dziwią badania w tym względzie prowadzone wobec robotów humanoidalnych. Z tym zaś wiąże się inny z ważnych tematów ciekawie przedstawianych i analizowanych w dysertacji, mianowicie kwestie przetwarzania języka naturalnego i związane z tym wyzwania. Wydaje mi się, że sposób w jaki Doktorant rozważa ten wątek świadczy o jego dojrzałości naukowej i trafności osądów.

To dziś jest bowiem jednym z najważniejszych tematów, a przywoływana przez Autora literatura świadczy, że doskonale orientuje się w najnowszych trendach.

W ramach swego namysłu nad roboetyką, Doktorant zatrzymuje się nad naszymi relacjami jako ludzi z robotami, przez co wywołuje temat „społecznej osobowości robotów", zwłaszcza ich traktowania. Wszystko to w kontekście tzw. postulatów de-antropocentryzacji i budowanej na tym haśle humanistyki, gdzie także i tu dał wyraz Doktorant swojej dojrzałości badawczej.

W ramach dialogu z Autorem chciałbym postawić trzy pytania.

1/ czy Doktorant uważa za możliwe odniesienie do robotów teorii cnót, która stanowi jak wiadomo jedną z najbardziej rozwijających się nurtów etyki?

2/ W ramach świętowania 500-lecia reformacji w 2017 roku przygotowano roboto-księdza, który udziela błogosławieństwa. Docierają również informacje dot. odprawiania celebracji religijnych przez roboty (np. w Japonii). Jak ocenia Autor te formy „socjalizacji" robotów?

3/ Transhumanizm – właściwie nie został szczegółowo omówiony, choć są odniesienia do tego nurtu w dysertacji: cyborgizacja człowieka a kwestia społecznej równości: czy nie prowadzi to wielkich nierówności i nowych problemów społecznych? Jakie problemy związane z podmiotowością robotów Doktorant ocenia jako najistotniejsze?

4. Wniosek końcowy

Po zapoznaniu się z treścią rozprawy doktorskiej Łukasza Sarowskiego, stwierdzam, że spełnia ona wszystkie wymogi formalne i stawiam wniosek do Rady Naukowej Dyscypliny Filozofia Katolickiego Uniwersytetu Lubelskiego im. Jana Pawła II w Lublinie o dopuszczenie do kolejnych etapów przewodu doktorskiego.