dr hab. Michał Parzuchowski, prof. Uczelni
Wydział Psychologii w Sopocie
SWPS Uniwersytet Humanistycznospołeczny
ul. Polna 16/20
81-745 Sopot

Gdańsk, 10 lipca 2023

Recenzja rozprawy doktorskiej mgr Katarzyny Zerki 'Intentionality attribution to social robots'

Tematyka rozprawy mgr Katarzyny Zerki osadzona jest w poznawczej psychologii społecznej i dotyczy spostrzeganej podmiotowości robotów społecznych. Główną tezą rozprawy jest znalezienie odpowiedzi na pytanie, jaką rolę w interakcjach ludzi z robotami mają oczekiwania ludzi wobec umysłowości algorytmicznych maszyn. Czy mechanizmy postrzegania społecznego właściwe dla interakcji między ludźmi (przypisywanie celowości obserwowanym zachowaniom), wykorzystywane są również przez obserwatorów zachowań humanoidalnych robotów? Czy obserwatorzy będą dostrzegać celowość działania niezależnie od tego czy klasyfikowane zachowania były wykonywane przez roboty o różnym stopniu antropomorfizacji?

Pani mgr Katarzyna Zerka wykonała cykl trzech badań poświęconych ustaleniu społeczno-poznawczych predyktorów (m.in. postaw wobec robotów, przekonań o unikalności ludzkiej natury, empatii czy antropomorfizacji) przewidujących tendencje w dostrzeganiu intencji w ambiwalentnych zachowaniach ludzi i robotów w zależności od warunków wydawania oszacowania (w warunkach presji czasu oraz po uprzedniej antropomorfizacji robotów). Jest to wyzwanie interesujące zarówno pod względem praktycznym, jak i szczególnie ciekawe środowisko do testowania nowych pomysłów badawczych z perspektywy teoretycznej.

Pod względem praktycznym rozprawa dotyczy sfery naszej aktywności, która w przyszłości ma szanse stać się dla nas codziennością. W obliczu dynamicznych prac korporacji informatycznych nad implementacją modeli językowych sztucznej inteligencji dylemat dostrzegania intencji w zachowaniach autorstwa robotów będzie zyskiwał na znaczeniu. Pod względem teoretycznym badane zjawisko jest również szalenie interesujące. Przypisywanie intencji sprawcy zaobserwowanego zachowania jest istotnym predyktorem dla szeregu podstawowych zmiennych funkcjonowania społecznego ludzi (np. przypisywania moralnej odpowiedzialności lub sprawności umysłu sprawcy por. Alicke, 2000; Malle i in., 2014, Weiner, 1995). Błędna atrybucja intencji jest przejawem optymalizacji kategoryzacji bodźców. Jest świadectwem przetargu między trafnością a wysiłkiem poznawczym zachodzącym w umyśle odbiorcy - błędne doszukiwanie się intencji (jako domyślny proces) sygnalizuje optymalne zużycie zasobów poznawczych potrzebnych do podjęcia podobnej decyzji. Błąd atrybucji intencji pozwala zatem uniknąć wrażenia niepewności/losowości dostarczając maksymalnego sensu/znaczenia przy niskim wysiłku poznawczym (Rosch, 1978). Tę tendencję poznawczą powinna szczególnie nasilać również wykorzystana w badaniach presja czasu. Badania mgr Zerki przybliżają nas do zrozumienia mechanizmów interakcji człowieka z robotami, choć zanim będzie można wykorzystać te interwencje w np. projektowaniu interfejsów, należałoby wypełnić istotne luki (które staram się wyliczyć w dalszej części recenzji).

Wyniki badań mgr Zerki sugerują, że postrzeganie robotów jako wolnych od błędów może prowadzić do przypisywania większej celowości ich zachowaniom. Wyniki nie w pełni potwierdziły przewidywania doktorantki dotyczące wzrostu atrybucji intencji w warunkach presji czasu, choć w przypadku zachowań prototypowo przypadkowych, oceny były zgodne z przewidywaniami modelu (klasyfikacja zachowań jako celowych była częstsza w warunkach presji czasu). W badaniu drugim manipulacja primingowa zadziałała odwrotnie do spodziewanego kierunku (nie stwierdzono większych różnic między robotami o wysokim i niskim poziomie cech antropomorficznych).

Niniejsza praca doktorska zawiera wyniki trzech badań (w tym jednego badania pilotażowego) uporządkowanych w logiczną całość. W sposób zwięzły, choć niepozbawiony wad, mgr Zerka raportuje kluczowe informacje o przebiegu i wyników swoich badań. Prostym i zrozumiałym językiem opisuje w wyczerpujący sposób wprowadzenie do stawianych hipotez. Rozprawa składa się z trzech części: teoretycznej, empirycznej i dyskusji uzyskanych wyników[1]. Podział tych treści na rozdziały nie jest jednak wystarczająco intuicyjny. Z jednej strony wprowadzenie teoretyczne skutecznie odpowiada na zasadne pytanie: czy staramy się ustalić stanu umysłu robota (*o czym teraz myśli?*), czy raczej rozpatrujemy zachowania robota wyłącznie poprzez odniesienie się do zaprojektowanego mechanizmu? Czy wyobrażenie sobie programowanej maszyny z symulowanymi cechami ludzkich zachowań (np. ekspresja mimiczna lub demonstrowanie ludzkiego opóźnienia w reakcji) wystarczy, aby humanoidalne roboty były postrzegane jako „posiadacze procesów umysłowych”? Czy prymowanie niskim vs. wysokim poziomem cech antropomorficznych robota wpływa na przypisywaną mu intencjonalność? Z drugiej strony jednak, nie rozumiem logiki kolejności zmiennych zaprezentowanych we wprowadzeniu. Czytelnik nie potrzebuje drugiego rozdziału (Human-Robot Interaction - w przeprowadzonych badaniach interakcja polegała jedynie na przyglądaniu się różnej wersji zdjęcia robota lub człowieka) dla zrozumienia propozycji testowanego w pracy modelu teoretycznego. Jednocześnie brakuje mi szerszego odniesienia planowanych badań oraz uzyskanych wyników do dyskusji o naturze intencjonalności czy szerzej podmiotowości z perspektywy psychologicznej (dostrzeganie intencji nawet w losowych zdarzeniach, intencje a budowanie teorii umysłu, celowość działań a budowanie postaw), kognitywistycznej czy filozoficznej.

Konkludując, zamiast licznych odwołań do interakcji człowiek-maszyna i użyteczności tego procesu (UX) (ta część nie jest elementem projektu empirycznego, więc zamiast do wprowadzenia nadaje się do ewentualnej pogłębionej dyskusji) oczekiwałem pogłębionych rozważań nad doszukiwaniem się znaczenia psychologicznego procesu atrybucji intencji i mechanizmu poznawczego modułu generowania podmiotowości u ludzi i robotów humanoidalnych. Przykładowo starałem się odnieść wyniki badań tego projektu do kartezjańskiego mitu “ducha w maszynie”, które wydaje się dobrym punktem wyjścia do rozważań o umysłowości robotów (intencjonalność robota może być przecież zaprogramowanym procesem a niekoniecznie sumarycznym oszacowaniem jego stanów). Argumentacja *intentional stance* Dennetta (1987), przywołana jest wyłącznie w kontekście doszukiwania się w zachowaniach robotów mechanizmu lub umysłu, podczas gdy sam Dennett w późniejszych pracach (np. 2015) sporo uwagi poświęca innym elementom tego

[1] Bardzo doceniam wysiłki mgr Zerki w zapewnieniu transparentości tego projektu - umieszczenie danych z wszystkich trzech badań oraz wykonanych analiz statystycznych w publicznym repozytorium.

procesu np. skali analizy. Dennett zauważa, że przyglądanie się organizmowi żywemu na poziomie komórkowym (mechanistyczne eksplorowanie najlepszych połączeń przez komórki nerwowe) przypomina pętle algorytmiczne sztucznej inteligencji czy zachowanie nanorobotów (powtarzając echa osiemnastowiecznej debaty nad paradoksem młynu Leibniza).

Dostrzegam kilka problemów z wnioskami z badań zawartymi w pracy, mam też kilka uwag do raportowanych wyników badań eksperymentalnych. Niemniej moje uwagi krytyczne mają za zadanie pomóc w przygotowaniu tekstu publikacji, jaki powinien się ukazać na bazie tej pracy i nie zmieniają pozytywnej oceny nowatorstwa i doniosłości tego cyklu badań. Szczegółowo omawiam moje uwagi poniżej.

1) dziwi mnie nieco, że pomimo słabości wykorzystanego pomiaru zmiennej zależnej (kiepska rzetelność kolejnych subskal) mgr Zerka nie zdecydowała się w kolejnych badaniach nieco inaczej uszczegółowić swój problem badawczy oraz ewentualnie rozszerzyć badany zakres projektu o ustalenie źródeł czy mechanizmu pośredniczącego doszukiwania się celowości o np. poprzez oszacowania dot. spójności celów (czy cele pokrywają się z zamierzeniami podmiotu) oraz ewentualną logiczność zachowania (czy zamierzenia są logiczną konsekwencją podjętego przez podmiot celu).

2) poszukiwania podmiotowości zachowań robotów mgr Zerka zawęziła do pomiaru proporcji klasyfikacji zachowań jako celowych różniących się wyjściowym wskaźnikami prawdopodobieństwa intencji wykonywanych przez roboty w wersji nisko i wysoko antropomorfizowanej (w porównaniu do warunku, w którym sprawcą zachowania był człowiek). W mojej ocenie projekt ten jest nie tyle punktem wyjścia do ustalenia zasad interakcji z robotami społecznymi, co szerzej świetnym początkiem dyskusji nad dostrzeganiem procesów umysłowych u prawdziwie inteligentnych robotów. W mojej ocenie w kolejnych krokach warto byłoby również ustalić sekwencję kolejnych przejawów podmiotowości i procesów epistemicznych skutecznych algorytmów robotów społecznych, takich jak m.in. postrzeganej kontrolowalności procesów umysłowych (por. Cusimano i Goodwin, 2019), aktywności percepcyjnej, tj. korygowania sygnału w oparciu o kontekst wydawanego sądu, analizę syntaktyczną, efektywność przeszukiwania pamięci, oceny planów czy w końcu doksastycznej stabilności przekonań, preferencji czy postaw.

3) brakuje analizy psychologicznego procesu uwikłanego w proces atrybucji intencji - analiza korelacji między zmiennymi objaśnianymi a kowariantami została wykonana (zamieszczona na OSF), lecz nie została opisana i przedyskutowana w pracy

4) Doktorantka nie podejmuje też dyskusji na temat niskiej mocy zrealizowanych badań eksperymentalnych (mała liczba badanych w porównywanych grupach) oraz braku analizy mocy a priori i braku prerejestracji. W badaniu pierwszym Doktorantka przebadała 206 osób, zaś w drugim 235. Przy zastosowanej liczbie pomiarów, kowariantów i liczbie zmiennych niezależnych (oraz potencjale możliwych interakcji zmiennych niezależnych z kowariantami), ograniczenie do przebadania po 28-38

osób w celkach wydaje się niewystarczające[2]. W kolejnym kroku przydałoby się przeprowadzenie analizy mocy i przebadanie dużo większej grupy badanych. Interpretacje w kategoriach “close to significance” nie są dobrym dowodem dla hipotez. Zamiast spekulować o kształcie wyników z małej próby, wymagana jest analizy mocy. Do przygotowywanej publikacji należałoby przeprowadzić replikację na wystarczająco mocnej próbie, aby móc wnioskować o sile efektów z braku różnic (analizy bayesowskie).

5) Mgr Zerka udostępniła bazy danych oraz pliki źródłowe analiz wykonane w JAMOVI i R, jednak opis analiz przedstawionych w rozprawie w mojej ocenie nie pozwala w pełni zorientować się w uzyskanych wynikach. Dla przykładu, skłonność do antropomorfizacji, empatia, negatywne postawy wobec robotów i przekonanie o wyjątkowości ludzkiej natury zostały uwzględnione jako kowarianty w analizie ANCOVA. W pracy zabrakło jednak bardziej podstawowego podsumowania uzyskanych wyników w postaci tabelarycznej prezentacji podstawowego modelu uzyskanych statystyk opisowych dla zachowań prototypowo przypadkowych, prototypowo intencjonalnych oraz zdarzeniach kontrolnych w podziale na zastosowane manipulacje) dla zmiennych objaśnianych (M, SD) wraz z korelacjami lub współczynnikami bety w hierarchicznej regresji liniowej dla badanych predyktorów. Analizę dodatkowo utrudnia fakt, że mgr Zerka nadmiernie używa skrótowców operacjonalizacji (nazwy skali np. NATIR czy BHNU) zamiast zmiennych na poziomie teoretycznym (postaw wobec robotów czy przekonań o unikalności natury ludzkiej).

6) W obliczu trudności z uśrednianiem wyników podskal (oraz znacznie różnej liczebności egzemplarzy w każdej z kategorii), czy krótkie opisy zachowań o różnym wskaźniku intencjonalności w ocenianych zachowaniach (zmienna zależna) można w procedurze sędziów kompetentnych ocenić na ciągłej skali intencjonalności (od prototypowo celowych (+1) przez zachowania ambiwalentne do prototypowo przypadkowych (-1) i potraktować jako kolejny predyktor lub kowariant w modelu?

7) Przeprowadzone badania nie zawierają kontroli skuteczności manipulacji (np. na koniec procedury powinno paść pytanie o to czy osoby badane zapamiętały kogo oceniały, lub czy domyślały się intencji działania sprawcy w kategoriach algorytmu czy procesów umysłowych). Nie ma też sprawdzenia skuteczności manipulacji presją czasu czy zastosowanym primingiem (czy badani rozwiązywali je krócej w warunku presji czasu; czy dostrzegli lub zinterpretowali zgodnie z intencją autorki wyświetlane filmy prezentujące robota gestykulującego vs. rzeźbiącego). Analiza nie zawiera również kontroli uważności badanych (np. poprzez uwikłanie pytań kontrolnych wśród pomiarów kowariantów np. “*W odpowiedzi na to pytanie odpowiedz “Nie wiem*”). Selektywne wypadanie z grup (problem z randomizacją do warunków badania - czy analizowano czas rozwiązywania badania w celkach?) i niskie wartości analizy rzetelności sugerują zróżnicowaną trudność badania w różnych warunkach eksperymentów (luki w danych sugerują, że badani w przypadku doświadczania

[2] przy okazji integracja warunków badania 1 i 2 pozostaje niezrozumiała (w tabeli 10 warunek 7 powinien mieć 34 osoby z badania 1 a raportowane jest 32).

trudności rezygnowali z odpowiedzi i mogli nie być zmotywowani do dalszego udziału w nużącym badaniu).

Pomimo powyższych uwag polemicznych i krytycznych, w mojej ocenie, praca wnosi znaczące i nowe ustalenia w obszarze psychologii społecznej - pozwala rozwijać nową metodologię przyszłych badań nad postrzeganiem podmiotowości robotów społecznych i jest znaczącą inspiracją do dalszych prób empirycznych. Badania wykonane przez Doktorantkę stanowią zatem znaczące uzupełnienie dotychczasowej wiedzy. Zrealizowany projekt empiryczny, mimo wspomnianych uchybień, dowodzi kompetencji naukowych Doktorantki, wskazując na umiejętność projektowania serii badań empirycznych, analizy ich wyników, formułowania wniosków i ograniczeń z nich płynących. Odkrycie to poszerza aktualną wiedzę i spełnia wymogi stawiane rozprawom doktorskim. Przedłożona do recenzji rozprawa doktorska spełnia zatem wymogi ustawowe (art. 13 ustawy z dnia 14 marca 2003 roku) związane z nadawaniem stopni i tytułów naukowych. Wnioskuję o dopuszczenie mgr Katarzyny Zerki do dalszych etapów przewodu doktorskiego.

Michał Parzuchowski