dr hab. Marcin Miłkowski, prof. nadzw. IFiS PAN

Instytut Filozofii i Socjologii PAN

**RECENZJA**

**pracy doktorskiej mgra Marcina Cichosza**

**pt. „Zintegrowany model złożonych działań intencjonalnych”**

## Wybór tematu i jego ujęcie

Rozprawa doktorska mgra Marcina Cichosza proponuje oryginalny, zintegrowany model złożonych działań intencjonalnych. Model ten, o charakterze teoretycznej propozycji badawczej, odwołuje się przede wszystkim do (1) filozoficznych rozważań na temat intencjonalności, (2) badań z zakresu psychologii intencji oraz (3) obliczeniowego modelowania uczenia się ze wzmacnianiem.

Praca opiera się na oryginalnej analizie teoretycznych podstaw bardzo ważnego zagadnienia w naukach o procesach poznawczych. Autor krytycznie odnosi się do dotychczasowych ujęć, traktując je jednak jako pomocne, wcześniejsze idealizacje, które można konkretyzować. Jest to szczególnie widoczne w jego podejściu do koncepcji intencjonalności J. Searle’a, która stanowi punkt wyjścia rozważań, lecz bynajmniej nie punkt dojścia, gdyż mgr Cichosz odrzuca i krytykuje większość przyjmowanych przez Searle’a kontrowersyjnych idei. Podobnie krytyczne nastawienie ma mgr Cichosz do stanowisk z psychologii intencji. Wreszcie modele uczenia się ze wzmacnianiem zostają uznane za obiecujące, lecz niewystarczające i zostają dopełnione przez modele oparte na kodowaniu predykcyjnym w odniesieniu do działań tzw. wyższego rzędu, a także przez opracowaną przez doktoranta koncepcją planowania opartego na sieci stanów intencjonalnych. Rozprawa nawiązuje więc do badań z wielu dyscyplin i z pewnością stanowi wyraźny przykład interdyscyplinarnych powiązań w kognitywistyce.

Nie trzeba dodawać, że doktorant podjął zagadnienie o kluczowym wręcz znaczeniu dla rozumienia, czym jest działanie intencjonalne, starając się precyzyjnie nakreślić obliczeniową architekturę odpowiedzialną za sterowanie takimi działaniami. Chodzi przy tym jednocześnie o działania stanowiące realizację naszych dalekosiężnych planów, jak i o najprostsze czynności motoryczne. W literaturze przedmiotu istotnie nie istnieje szczegółowy model działania o tak

szerokim zakresie, mimo że wielu badaczy dąży do unifikacji teorii działania (i poznania). W modelu doktoranta odbicie znajdują wcześniejsze koncepcje psychologiczne, filozoficzne i wywodzące się z obliczeniowego modelowania w neuronauce, które zostały zintegrowane w sposób interesujący, nowatorski i ambitny.

Wybór tematu uznaję więc za trafny, a cel – opracowanie nowego modelu złożonych działań intencjonalnych – za dobrze zrealizowany.

**Struktura i forma pracy**

Praca mgra Cichosza liczy mniej więcej 20,7 arkuszy wydawniczych (ok. 460 stron maszynopisu standardowego), jest więc stosunkowo obszerna. Składa się z pięciu rozdziałów oraz zakończenia i bibliografii. Praca jest niestety pozbawiona indeksów (tematycznego i autorskiego).

Forma językowa rozprawy pozostawia nieco do życzenia. Dosyć liczne są literówki (nawet w podziękowaniach, gdzie widnieje „Kognitywistki" zamiast „Kognitywistyki"), zwłaszcza w terminologii anglojęzycznej („sens of agency" i „sens of urge", s. 10 i 142) i nazwiskach (np. „Dayen" zamiast „Dayan", s. 99, „Markova" zamiast „Markowa", s. 130). Nie wiem, dlaczego Richard Sutton na s. 218 stał się „Bartonem Suttonem". Usterką podobnego rodzaju jest przypisanie Robertowi A. Rescorli i Allanowi R. Wagnerowi cechy bycia „niemieckimi badaczami" (w istocie pracowali oni na Yale University; Rescorla urodził się w Filadelfii, a Wagner w Springfield).

Na s. 76 pozostawiono anglojęzyczny komunikat o błędzie dotyczącym lokalizacji (numeru strony) Diagramu 3, znajdującego się istotnie wcześniej na s. 65. Na s. 102 wzór opisujący funkcję wartości w uczeniu się ze wzmacnianiem nie występuje tam, gdzie powinien, a więc po pierwszym akapicie, lecz w zupełnie nieoczekiwanym miejscu, niżej, w środku długiego cytatu.

Mam też zastrzeżenia dotyczące terminologii – doktorant posługuje się terminem „prior intencja", który ma charakter dziwnego angielsko-polskiego zlepka. Nie podaje jednak żadnego uzasadnienia, dlaczego nie można w tym miejscu zastosować po prostu terminu „intencja uprzednia", który nie budziłby np. wątpliwości dotyczących wymowy tego terminu (czy należy stosować poprawną wymowę angielską, tj. /praɪəʳ/, czy też go spolszczać?). Sporadycznie zdarzają się też błędy w tłumaczeniach cytatów (które w oryginalnej wersji mgr Cichosz przytacza w całości w przypisach dolnych), np. na s. 76 czytamy, że filozofowie „ciągle próbują nadać sens przyczynowości umysłowej", podczas gdy lepiej byłoby napisać, że ciągle starają się zrozumieć przyczynowość umysłową (tak należy oddawać wyrażenie „make sense of" w tym kontekście). Niepotrzebne jest też wprowadzanie neologizmu „preplanowanie" jako odpowiednika (co prawda, rzadkiego, lecz notowanego w angielskich słownikach) wyrazu „preplanning" (s. 147), gdzie wystarczyłoby po

prostu „wcześniejsze zaplanowanie".

Innego rodzaju wadą rozprawy jest wprowadzanie bardzo wielu skrótowców (chociażby „ZMDI" na oznaczenie modelu bronionego w rozprawie), które czasami używane są co najwyżej parę razy, przez co jedynie utrudniają czytelnikowi lekturę i sprawiają wrażenie symbolomanii. Skądinąd jednak mgr Cichosz nie epatuje zbyt skomplikowanymi formalizmami, w bardzo prosty i zrozumiały sposób przedstawiając niekiedy dosyć złożone matematycznie koncepcje, takie jak kodowanie predykcyjne czy aktywne wnioskowanie K. Fristona. Samo to stanowi powód do chwały, bo nazbyt rzadko zdarza się w literaturze przedmiotu.

Do pozostałych usterek pracy można zaliczyć niekompletną bibliografię, np. brak tam pozycji Braun i in. 2018, Ciechanowski 2015, Colombo 2014. Dosyć ryzykowne jest też cytowanie wypowiedzi z prezentacji wykładowych (to dotyczy np. cytatu z mojego wykładu kursowego), gdy można wtedy dosyć mocno pobłądzić. I tak mgr Cichosz przypisuje mi pogląd, iż do najbardziej popularnych ram teoretycznych w kognitywistyce należą „koncepcja rozszerzonego umysłu, enaktywizm, predykcyjna teoria umysłu, teleosemantyka" (s. 31), podczas gdy na wykładzie z zaawansowanych zagadnień filozofii umysłu, podsumowującym cały semestr, wymieniałem podstawowe *zagadnienia filozoficzne* omawiane w tym kursie. Nie sądzę, by koncepcja rozszerzonego umysłu była popularną ramą teoretyczną w samej kognitywistyce, bo do tej pory nie odegrała ona szczególnie istotnej roli w praktyce eksperymentalnej i stanowi tylko pewnego rodzaju metafizyczną koncepcję umysłu, a enaktywizm na pewno nie jest równie popularny, jak komputacjonizm.

Mam też pewne zastrzeżenia co do stosowanej symboliki. Niestety, tak jak wielu innych autorów, mgr Cichosz posługuje się symbolami niealfabetycznymi bez wprowadzania ich definicji. Dotyczy to np. strzałek w wyrażeniach typu „umysł → świat" i czytelnik musi się domyślać, o jakiego rodzaju relację chodzi: czy umysł dopasowuje się do świata, czy też świat do umysłu (np. s. 56). Z kontekstu można domyślić się, że „→" oznacza relację „bycia uzgadnianym z", ale powinno zostać to zdefiniowane wprost (taka sama strzałka jednak prawdopodobnie oznacza oddziaływanie przyczynowe na s. 156, wers 6 od dołu). Podobnie niejasna jest semantyczna interpretacja różnych średnic okręgów oznaczających reprezentacje (Diagram 6, s. 126); komentarz słowny w tekście głównym kończy się znakiem zapytania i pozostaje niezrozumiały.

Przy wszystkich tych zastrzeżeniach podkreślić należy jednak, że wywód przeprowadzony jest starannie i klarownie, z dużą intelektualną dyscypliną i precyzją, a jednocześnie koncepcje teoretyczne omawianych autorów są interpretowane rzetelnie i z dochowaniem zasady życzliwości. Nieprecyzyjne wypowiedzi są bardzo nieliczne (np. na s. 55 można się dowiedzieć, że K. Bielecka

broniła znaturalizowanych ujęć reprezentacji umysłowych [co jest prawdą], „a w szczególności tych, które mają charakter intencjonalny" – cóż, w książce Bieleckiej [2019] przyjęto założenie, że niezbędną cechą reprezentacji jest intencjonalność, więc innymi zajmować się z konieczności nie mogła).

**Zakres i treść pracy**

Praca składa się z pięciu rozdziałów, uporządkowanych w kolejności, w jakiej poszczególne koncepcje są wykorzystywane w autorskim zintegrowanym modelu działań intencjonalnych. We wprowadzeniu doktorant wprowadza elementarne pojęcia, definiuje problem podejmowany w rozprawie, a także broni podejścia multidyscyplinarnego. Określa tam też tezę rozprawy: „Złożone działanie intencjonalne to system dwóch współdziałających mechanizmów: (i) uczenia się ze wzmacnianiem oraz (ii) planowania na podstawie wiedzy w formie sieci stanów intencjonalnych" (s. 33). Następnie (s. 35) wskazuje, że celem rozprawy jest określenie relacji między (i) a (ii). Przedstawia wreszcie zarys zintegrowanego modelu działań intencjonalnych oraz plan pracy.

W rozdziale 2 doktorant rekonstruuje koncepcję intencjonalności amerykańskiego filozofa Johna Searle'a, gdyż – zdaniem doktoranta – „zawiera ona szereg cennych spostrzeżeń i propozycji" (s. 52). Zaletą tą ma być uwzględnienie „wiedzy empirycznej o działaniu układu nerwowego oraz biologicznych podstawach zachowań ludzkich" (s. 53). Jak się jednak znacznie później okazuje, doktorant w istocie uznaje, że Searle uwzględnia tę wiedzę wyłącznie w sferze deklaracji, a faktycznie całkowicie ją ignoruje; jest to też powodem dosyć wyrazistej krytyki w późniejszych rozdziałach rozprawy. Do najważniejszych przywoływanych składników koncepcji Searle'a należą:

- założenie o powiązaniu intencjonalności ze świadomością;

- założenie, że tzw. tło i wiedza-jak nie ma charakteru reprezentacyjnego;

- założenie, że w skład tzw. tła wchodzą dyspozycje biologiczne i lokalne praktyki kulturowe (tworzące zbiory rozłączne);

- schemat przebiegu działania intencjonalnego, od deliberacji, przez uformowanie intencji uprzedniej, przez działanie – obejmujące też pojawienie się intencji w działaniu – po sam fizyczny ruch ciała

- założenie o przyczynowym charakterze intencji (a w szczególności świadomego przeżycia towarzyszącego intencji)

- wreszcie tezy o nieeliminowalności intencjonalności i biologicznym charakterze.

W kolejnych rozdziałach niemal wszystkie te założenia podlegają systematycznemu podważeniu (i są one, w opinii niżej podpisanego, rzeczywiście fałszywe; do sprawy wrócę).

W rozdziale 3 doktorant opisuje koncepcję uczenia ze wzmacnianiem jako koncepcję działania intencjonalnego. Wiąże też uczenie ze wzmacnianiem z płodną w neuronauce hipotezą, że błąd predykcji nagrody w uczeniu ze wzmacnianiem ma charakter dopaminergiczny. Mgr Cichosz przedstawia algorytmy uczenia ze wzmacnianiem jako modele obliczeniowe ludzkiego działania (w sposób dosyć przystępny, np. posługując się diagramami, por. Diagram 5, s. 101 – chociaż tam też trochę brakuje oznaczeń relacji symbolizowanych strzałkami, ale także w postaci pseudokodu, s. 103). Prezentacja jest dosyć szczegółowa, może nawet posunięta do pewnej pedanterii, kiedy prezentowane są stany agenta w zależności od rozkładu nagród, zmieniające się w czasie (s. 104-105). Szczególnie istotną rolę w pracy odgrywa algorytm uczenia się ze wzmacnianiem oparty na różnicy czasowej.

Należy odnotować, że na s. 113 doktorant podkreśla, że klasyczne implementacje algorytmu uczenia się ze wzmacnianiem nie pozwalają reprezentować sieci stanów intencjonalnych, gdyż nie mogą wykorzystywać wiedzy w formie symbolicznej. Z tego wyciąga wniosek, że nie jest adekwatnym modelem złożonych działań intencjonalnych, lecz dalej wskazuje możliwe kierunki rozwoju modelowania, które mogłyby do tego posłużyć.

W rozdziale 4 doktorant opisuje wyniki badań z zakresu psychologii intencji, wskazujące na interpretacyjny (a nie przyczynowy, wbrew Searle'owi) status stanów intencjonalnych towarzyszących działaniom intencjonalnym. Eksperymenty te ograniczają się do prostych działań, stąd też wyniki nie mogą być swobodnie ekstrapolowane na działania złożone. W rozdziale tym omawiane są wyniki klasycznych eksperymentów B. Libeta. Doktorant krytycznie podchodzi do założeń metodologicznych jego eksperymentów oraz do hipotetycznego mechanizmu weta. Wskazuje jednak, że dominującym poglądem wśród psychologów intencji, takich jak P. Haggard, jest uznanie, że intencja w działaniu (w sensie Searle'a) ma status „świadomościowego korelatu procesów przygotowujących ruch" (s. 151). W rozdziale tym krytycznie omawiane są też koncepcje Daniela Wegnera, zwłaszcza dotyczące poczucia sprawstwa. W świetle wyników badań poczucia sprawstwa doktorant stwierdza, że „intencja w działaniu okazała się zbyt prostym konstruktem teoretycznym" (s. 175). W zamian doktorant proponuje znacznie bardziej złożony model (Diagram 8, s. 176), oparty na modelu sterowania motoryką Wolperta-Mialla, a więc uwzględniającym kluczową rolę modeli wyprzedzających. W modelu tym intencja w działaniu zostanie przypisana do instrukcji ruchowej, a intencja uprzednia – do mechanizmu planującego ruch (wydającego

instrukcje ruchowe). Poczucie sprawstwa jest z kolei powiązane z modelem wyprzedzającym i kopią eferentną. Następnie mgr Cichosz doprecyzowuje funkcję intencji w działaniu i funkcję poczucia sprawstwa.

Rozdział 4 zostaje podsumowany stwierdzeniem, że psychologowie intencji koncentrują się przede wszystkim na działaniach prostych, a złożone rozumieją wyłącznie jako sekwencje działań prostych (s. 187). Zdaniem doktoranta jest to oparte „na wątłych podstawach".

Rozdział 5 przynosi oryginalną syntezę – model integrujący elementy koncepcji przedstawionych w poprzednich rozdziałach. Mgr Cichosz wskazuje tam na bardzo liczne niedostatki konceptualne w podejściu Searle'a, tzw. naturalizmu biologicznego: (1) spekulatywny charakter tej koncepcji; (2) uznanie opisu zjawiska za jego wyjaśnienie; (3) nietrafna charakterystyka wielopoziomowych podejść badawczych" (s. 193). Doktorant krytykuje kolejno różne składowe koncepcji Searle'a, łącznie z mało precyzyjnym określeniem zarówno zakresu, jak i charakteru tzw. tła. Następnie przechodzi do charakterystyki pojęcia działania intencjonalnego. Otóż wg doktoranta „działanie agenta jest intencjonalne wtedy i tylko wtedy, gdy:

(a) jest ono wykonalne na gruncie wiedzy agenta o nim samym i o jego własnym położeniu,
(b) zaplanowany jako skutek działania stan zamierzony jest bardziej wartościowy dla agenta niż stan zastany,
(c) agent chce poprawić swoje położenie poprzez podjęcie działania" (s. 205).

Kolejne części rozdziału przynoszą kolejne, coraz bardziej złożone zintegrowane modele działania intencjonalnego, analizowane na trzech etapach – (1) definiowania i analizy wymagań, (2) konstrukcji i implementacji systemu, (3) testowania (s. 208). Co prawda, doktorant stwierdza, że faza testowania ze względu na „teoretyczny charakter prowadzonych w pracy rozważań" została w dysertacji pominięta (s. 208, przyp. 97), to pozwolę sobie nie zgodzić się z tą uwagą. Otóż testowanie nie musi być rozumiane w sensie sprawdzania empirycznego – tzn. walidacji modeli w technicznym rozumieniu pojęcia walidacji (czyli porównania z wynikami eksperymentalnymi na odpowiednim poziomie abstrakcji). Testowanie obejmuje też tzw. weryfikację modelu, czyli sprawdzenie, czy model odpowiada samym wymaganiom, także tym, które wydają się oczywiste – i takie testowanie w dysertacji jest przeprowadzone przez przytaczanie przykładów działania, do którego opisu dana wersja modelu nie wystarcza (np. s 227 – „powyższy model nie wystarcza do opisu zachowań wyższych zwierząt").

Nie będę przedstawiał wszystkich elementów oryginalnej propozycji oraz wszystkich wersji modelu działania, ograniczając się do ich naszkicowania oraz konstatacji, że doktorant z wysoką świadomością teoretyczną wskazuje liczne idealizacyjne założenia w jego konstrukcji. I tak np.

zakłada, że struktura systemu realizującego działania jest zasadniczo niezmienna (co oznacza pominięcie faktu rozwoju poznawczego). Istotną różnicą między dwoma pierwszymi wersjami modelu a modelem 1.2 jest wykorzystanie koncepcji kodowania predykcyjnego K. Fristona do modelowania zachowań „wysokiego poziomu" (s. 240). Kodowanie predykcyjne bywa zresztą rozumiane jako koncepcja ogólniejsza od uczenia się przez wzmacnianie, dlatego też jest to dosyć naturalne posunięcie. W wersji 2.0 działanie intencjonalne ma być osadzone w sieci procesów poznawczych – pojawia się tam sieć stanów intencjonalnych (s. 254-255). Doktorant w przekonujący sposób odrzuca założenie Searle'a, że tło ma charakter niereprezentacyjny (s. 259). Wprowadza się pojęcie reprezentacji niezależnych od kontekstu, którym przypisuje się tzw. aspektową formę (s. 263). Ogólnie rzecz biorąc, w wersji 2.0 modelu doktorant stara się zintegrować koncepcję Searle'a z kodowaniem predykcyjnym, co nie zawsze jest łatwe; wskazuje np. że trudno wskazać „analogon holizmu znaczeniowego w propozycji Fristona" (s. 264). Kluczową nowością wprowadzaną w modelu 3.0 jest realizacja działań wg planu. Uwzględnia się też procesy i zmiany rozwojowe prowadzące do wytworzenia umiejętności deliberacji.

W zakończeniu rekapitulowana jest treść rozprawy i założenia modelu, który ma łączyć trzy podsystemy: (1) podsystemu hierarchicznego uczenia się ze wzmacnianiem z optymalizacją domenową; (2) podsystemu planowania i realizacji planów oraz (3) podsystemu zarządzania siecią stanów intencjonalnych.

**Uwagi krytyczne i polemiczne**

Ze względu na wagę poruszanego w dysertacji zagadnienia i samą jej objętość łatwo wysunąć wiele zastrzeżeń wobec proponowanych w niej rozwiązań czy samej jej konstrukcji. Zanim jednak do tego przystąpię, muszę podkreślić, że praca wyraźnie świadczy o dużej samodzielności naukowej doktoranta i stanowi oryginalne rozwiązanie postawionego w niej problemu. Istotnie integracja planowania z uczeniem się ze wzmacnianiem jest zagadnieniem trudnym i wymagającym pracy przede wszystkim teoretycznej. Oceniana rozprawa z pewnością może przyczynić się do lepszego rozumienia działań intencjonalnych.

Przystąpię teraz do wskazania kilku uwag krytycznych.

*Definicja działania.* Przedstawiona w rozprawie definicja działania jest niepoprawna, jeśli ma być rozumiana jako definicja sprawozdawcza (a nawet jako definicja regulująca jest problematyczna, bo nie pozwala uznać działań wybieranych w ramach homeostazy za właściwie wybrane). Problematyczne są warunki (b) i (c). Otóż podmiot działający nie zawsze musi działać tak, aby uzyskiwać bardziej wartościowy stan niż stan zastany. Nie mam tu na myśli możliwości działania wbrew własnemu interesowi (co można byłoby uznać za pewnego rodzaju zaburzenie normalnego

działania, np. często analizowane w filozofii zjawiska określane mianem „słabości woli" czy „akrazji", lecz zwykle są one uznawane za paradygmatyczne przykłady zaburzonej racjonalności), lecz prosty fakt, że czasem podmiot chce tylko utrzymania pożądanego stanu zastanego. Np. ja mogę chcieć pójść na spacer i w trakcie spaceru mogę nadal chcieć spacerować, chociaż spacer nie jest w danym momencie dla mnie bardziej wartościowy niż chwilę wcześniej. Warunek (b) nie powinien wymagać, by stan był bardziej wartościowy, lecz równie lub bardziej wartościowy (relacja bycia większym lub równym, $\geq$). Warunek (c) – analogicznie – podmiot działający nie musi poprawiać swojego położenia, wystarczy, że go nie pogorszy. Warto zauważyć, że doktorant podkreśla istotność homeostazy w regulacji organizmu, a działania podejmowane w jej ramach zwykle będą skupione na podtrzymaniu pożądanego stanu, a nie na jego poprawieniu (homeostaza podtrzyma nas przy życiu, ale nie zapewni nieśmiertelności).

*Koncepcja intencjonalności Searle'a.* Konstrukcja rozprawy utrudnia zrozumienie, co doktorant przyjmuje, a co odrzuca z koncepcji Searle'a. Lektura rozdziału 2 może sugerować, że wszystkie założenia są traktowane jako oczywiste i bezdyskusyjne; chyba znacznie lepiej byłoby od razu zaznaczyć, że będą one podważane. Jak już wskazywałem, założenie, że tzw. tło i wiedza-jak nie ma charakteru reprezentacyjnego, zostaje wyraźnie odrzucone, podobnie jak założenie o przebiegu działania intencjonalnego. Trzy elementy pozostają wszakże w pewnym zawieszeniu – nie wiadomo, co jest przyjmowane w modelu bronionym w doktoracie.

Po pierwsze, nie jest jasne, jaka jest rola świadomości w działaniu i jego strukturyzacji. Jest to tym ważniejsze, że np. struktura czynności językowych nie podlega pełnemu uświadomieniu: nikt w introspekcji nigdy nie jest w stanie zdać sprawy z procesów generowania gramatycznych zdań. A struktura takich działań – nieredukowalnych, co wiemy już od czasu pamiętnej recenzji Chomsky'ego (1959), do prostych sekwencji działań prostych – powinna być uwzględniona w wymaganiach wobec każdego rozsądnego modelu organizacji zachowania. Jest to truizm przynajmniej od czasów Lashleya (1951). Tymczasem Searle jest zmuszony, jak sądzę, uznać, że gramatyka jest potencjalnie uświadamialna.

Po drugie, założenie, że w skład tzw. tła wchodzą dyspozycje biologiczne i lokalne praktyki kulturowe, zdradza ewidentne uproszczenie – wiadomo, że praktyki kulturowe (takie jak posługiwanie się językiem naturalnym) wymagają istnienia dyspozycji biologicznych. W przypadku uniwersalnych zdolności ludzkich, jak komunikacja w języku naturalnym, trudno uznawać te umiejętności za „lokalne". Nie znajduję jasnego określenia, czy w modelu 3.0 to odróżnienie ma rację bytu; możliwe, że w ogóle należałoby je pominąć.

Po trzecie, nie jest jasne, co w istocie na temat istoty intencjonalności sądzi doktorant – czy jest ona

nieredukowalna? A jeśli jest nieredukowalna, to jak można sądzić, że ma charakter biologiczny? Wydaje się to dosyć wątpliwe. Co więcej, enigmatycznie brzmią dla mnie stwierdzenia mgr. Cichosza, że koncepcja Searle'a nie różni się od emergentyzmu, który został zdefiniowany jako stanowisko, „zgodnie z którym w złożonym systemie można wyróżnić jednostki niższego rzędu oraz wynikające z ich kompozycji jednostki wyższego rzędu". Jeśli ta definicja ma charakter sprawozdawczy, to jest wadliwa, bo emergentyzm okaże się też stanowiskiem bronionym przez najbardziej paradygmatycznych redukcjonistów, np. Putnama i Oppenheima (1958), którzy istotnie uważali, że istnieje pewna (wręcz czasoprzestrzenna) hierarchia kompozycyjna obiektów badanych przez naukę. Emergentyzm jest stanowiskiem zakładającym zwykle znacznie więcej, np. to, że organizacja jednostek wyższego rzędu nie daje się przewidzieć lub wyjaśnić w kategoriach jednostek niższego rzędu i relacji między nimi; można też uważać, że własności systemowe nie są prostymi agregatami własności jednostek niższego rzędu.

*Intencja w działaniu.* Muszę przyznać, że żywię ogromne wątpliwości co do istnienia bytu oznaczanego terminem „intencja w działaniu". Jak wskazuje Bence Nanay (2013), pojęcie „intencja w działaniu" służy tylko obronie filozoficznej koncepcji, zgodnie z której działanie opiera się na przekonaniach. Jest ono jednak w większości paradygmatycznych przykładów Searle'a – dotyczących tylko motoryki, jak jego słynny przykład ze wstaniem od komputera i chodzeniem po pokoju – zbędne, gdyż wystarczy w nich stwierdzić, że działanie wymaga tylko reprezentacji pragmatycznej, nie zaś (nieuświadomionych) przekonań, pragnień i uprzednich intencji. Czyż intencje w działaniu nie są po prostu epicyklami mającymi bronić filozoficznej koncepcji, zgodnie z którą działanie wymaga przekonań i pragnień? Badania psychologów intencji wskazują, że mają tak czy inaczej charakter epifenomenalny, więc może w dobrej teorii działania nie musi tych bytów w ogóle być.

*Uczenie się ze wzmacnianiem ma duże ograniczenia*. Dysertacja czerpie z zasłużenie wpływowej koncepcji uczenia się ze wzmacnianiem, która okazała się niesłychanie głośna zwłaszcza w odniesieniu do głębokich sieci neuropodobnych, uzyskujących mistrzowskie wyniki w grach itp. Doktorant przypisuje tej koncepcji liczne ograniczenia, w tym trudności zastosowania jej do reprezentacji symbolicznych. To ograniczenie wydaje się jednak trudność nieco na wyrost, biorąc pod uwagę wykorzystanie uczenia się przez wzmacnianie w klasycznych już pracach nad grami sygnalizacyjnymi (Skyrms 2010). Mówiąc w największym skrócie, można oczekiwać, że uczenie się ze wzmacnianiem może doprowadzić do powstawania komunikacji symbolicznej, a przy tym w sposób uznawany za biologicznie wiarygodny. Większa trudność jest wiązana w literaturze przedmiotu nie tyle z reprezentacjami symbolicznymi, ile z dosyć spektakularnymi porażkami tych metod uczenia w odniesieniu do głębokich sieci neuronowych. Było to przedmiotem ożywionej

debaty od dosyć znanego wpisu blogowego Aleksa Irpana (2018). Niestety, cała ta debata jest wielkim nieobecnym w recenzowanej rozprawie doktorskiej. A problemy są niebanalne: uczenie się ze wzmacnianiem jest dosyć mało efektywne, trudno jest wyuczyć funkcje nagrody, trudno unikać tzw. lokalnych optimów. W obecnym stanie rzeczy, jak powiada Irpan, bardzo trudno mieć nadzieję na skuteczne stosowanie tej metody w odniesieniu do 70% interesujących problemów.

Co więcej, można mieć wątpliwości, czy uczenie się ze wzmacnianiem pozwala w pełni zrozumieć złożone działania intencjonalne nie ze względu na trudności modelowania złożonych działań, lecz ze względu na nieadekwatny model motywacji. Jak pokazują liczne badania, zewnętrzne nagrody mają znacznie mniejszą rolę motywacyjną niż wewnętrzne – uzyskanie nagród daje zdecydowanie mniej efektów niż realizacja celu, jakim jest uczenie samo w sobie (Elliott i Dweck 1988). *Prima facie*, wydaje się jednak, że odróżnienie to nie znajduje odzwierciedlenia w teoriach uczenia ze wzmacnianiem. Być może jednak uwzględnienie literatury z zakresu teorii motywacji oraz osiągnięć mistrzowskich (Ericsson i Pool 2016) pomogłoby nakreślić model 4.0.

*Pojęcie planu a model TOTE*. Lektura rozdziału 4 nasuwa proste pytanie: czym w istocie różni się proponowany model od klasycznej koncepcji TOTE (Test-Operate-Test-Exit), która stanowiła przykład jednej z pierwszych teorii kognitywistycznych (Miller, Galanter i Pribram 1980)? Widać, że teoria ta musiałaby zostać wzbogacona o uczenie się, lecz mam wątpliwości, czy samo pojęcie planu odbiega od tego, co już proponowano w roku 1967.

*Format reprezentacji w sieci stanów intencjonalnych*. Nie jest jasne, jaki format mają reprezentacje wchodzące w skład sieci stanów intencjonalnych postulowanej przez doktoranta. Czy są to reprezentacje o postaci sądów? Czy są one z konieczności wyrażane w jakimś kodzie cechującym się składnią logiczną? We współczesnych modelach opartych na sieciach głębokich do czynienia mamy z sieciami stanów reprezentacyjnych, zwłaszcza w badaniach z zakresu przetwarzania języka naturalnego. Tylko że węzły tych sieci nie muszą być zawsze pełnymi zdaniami (czasem są to nawet reprezentacje subsymboliczne w przypadku bardzo skutecznych w tłumaczeniu maszynowym sieci na poziomie części wyrazów słownikowych, odpowiadających raczej składnikom morfologicznym niż leksemom – a czasem nawet są to sieci pojedynczych znaków). Rozprawa, niestety, nie pozwala łatwo rozstrzygnąć tej wątpliwości.

*Myślenie a kodowanie predykcyjne.* Ostatnia moja wątpliwość wiąże się z możliwością modelowania myślenia w kategoriach hierarchicznego kodowania predykcyjnego. Doktorant zakłada, że kodowanie predykcyjne pozwala zamodelować złożone działania intencjonalne, ale wydaje się, że właśnie tu tkwi słabość tego podejścia (Williams 2020; Williams 2018). Mówiąc krótko, kodowanie predykcyjne ma stosunkowo łatwą do określenia hierarchię przetwarzania w

przypadku procesów postrzeżeniowych, lecz hierarchia w przypadku procesów myślenia pozostaje niezdefiniowana, gdyż nie wiadomo, co wyznacza porządek w tej hierarchii. To w istocie pięta Achillesa hierarchicznego kodowania predykcyjnego, a to znaczy, że niespecjalnie może ułatwić modelowanie złożonych działań intencjonalnych. Trudność ta dotyczy nie tylko koncepcji mgra Cichosza, ale jest w ogóle nieprzezwyciężonym – być może na razie – problemem całej tradycji badawczej opartej na hierarchicznym kodowaniu predykcyjnym.

**Wnioski**

Recenzowana rozprawa świadczy o szerokiej erudycji doktoranta, ale także umiejętności integrowania koncepcji wywodzących się z różnorodnych tradycji badawczych. Podejście mgra Cichosza wykorzystuje osiągnięcia wielu dyscyplin, lecz rezultat nie jest bynajmniej mechanicznym zlepkiem wcześniejszych koncepcji, lecz starannie skonstruowanym szkicem modelu obliczeniowego. Jest to szkic, gdyż doktorant przedstawia tylko zarys modelu obliczeniowego, który mógłby zostać opracowany w przyszłości. Niemniej jednak zgodnie z utrwaloną już w kognitywistyce metodologią, której najgorętszym orędownikiem był David Marr, to właśnie taki zarys jest najważniejszym krokiem do pełnego wyjaśnienia badanego zjawiska. Zarys ten został przedstawiony przede wszystkim na poziomie obliczeniowym w sensie Marra, lecz zawiera też pewne wskazówki dotyczące poziomu algorytmów i reprezentacji (dzięki wskazaniu typów algorytmów – uczenie się ze wzmacnianiem i kodowanie predykcyjne – a także struktur danych – sieci stanów). Może więc posłużyć jako specyfikacja przyszłych prac z zakresu modelowania obliczeniowego. Z klasycznej perspektywy kognitywistycznej jest to więc robota wręcz wzorcowa, gdyż określa zarówno funkcjonalną charakterystykę badanego zjawiska, jak i jej możliwą obliczeniową strukturę.

Podjęty przez doktoranta temat należy do najtrudniejszych w badaniach nad poznaniem i działaniem. Nie sposób go wyczerpać w jednej rozprawie ani nawet w serii wydawniczej poświęconej temu zagadnieniu. Praca doktorska przedstawia jednak dobrą syntezę teoretyczną, która wnosi nową jakość do badań nad działaniem intencjonalnym.

W świetle powyższych uwag stwierdzam, że rozprawa mgra Marcina Cichosza **spełnia wymagania stawianych pracom doktorskim z zakresu nauki komunikacji społecznej i mediach. Wnioskuję o dopuszczenie go do dalszych etapów postępowania w przewodzie doktorskim**.

**Bibliografia**

Chomsky, Noam. 1959. *Review of Verbal Behavior by B. F. Skinner*. „Language” 35 (1): 26–58.
Elliott, Elaine S. i Carol S. Dweck. 1988. *Goals: An Approach to Motivation and Achievement.*

„Journal of Personality and Social Psychology" 54 (1): 5–12. doi:10.1037/0022-3514.54.1.5.
Ericsson, Karl Anders i Robert Pool. 2016. *Peak: Secrets from the New Science of Expertise*. London, The Bodley Head.
Irpan, Alex. 2018. *Deep Reinforcement Learning Doesn't Work Yet*. https://www.alexirpan.com/2018/02/14/rl-hard.html.
Lashley, Karl S. 1951. *The problem of serial order in behavior*. W „Cerebral Mechanisms in Behavior", red. L.A. Jeffries: 112–147. New York, John Wiley & Sons.
Miller, George A., Eugene Galanter i Karl H. Pribram. 1980. *Plany i struktura zachowania*. Tłum. Aldona Grzybowska i Adam Szewczyk. Warszawa, Państwowe Wydawnictwo Naukowe.
Nanay, Bence. 2013. *Between perception and action*. Oxford, Oxford University Press.
Oppenheim, Paul i Hilary Putnam. 1958. *Unity of science as a working hypothesis*. W „Concepts, theories, and the mind-body problem. Minnesota Studies in the Philosophy of Science. Vol. 2", red. Herbert Feigl, M. Scriven i G. Maxwell: 3–36. Minneapolis, Minnesota University Press.
Skyrms, Brian. 2010. *Signals: evolution, learning, & information*. Oxford, New York, Oxford University Press.
Williams, Daniel. 2018. *Hierarchical Bayesian Models of Delusion*. „Consciousness and Cognition" 61: 129–147. doi:10.1016/j.concog.2018.03.003.
Williams, Daniel. 2020. *Predictive Coding and Thought*. „Synthese" (197): 1749–1775. doi:10.1007/s11229-018-1768-x.

Warszawa, 10 lutego 2022

Podpisany elektronicznie przez
Marcin Miłkowski
10.02.2022
16:29:13 +01'00'

00-330 Warszawa
ul. Nowy Świat 72
tel.: +48 22 826 71 81
tel.: +48 22 828 32 03
fax: +48 22 826 99 48
secretar@ifispan.waw.pl
www.ifispan.pl