UNIWERSYTET im. ADAMA MICKIEWICZA

W POZNANIU

Wydział Psychologii i Kognitywistyki

Marcin Cichosz

Zintegrowany model złożonych działań intencjonalnych

Integrated model of complex intentional actions

Praca doktorska napisana pod kierunkiem

dra hab. Andrzeja Klawitera

POZNAŃ 2021

Autor pracy pragnie podziękować:

Promotorowi - za opiekę merytoryczną, za inspiracje, za nieustające wsparcie i cierpliwość okazaną podczas pisania pracy. Szczególnie wdzięczny jestem za przenikliwe i konstruktywne uwagi, których Profesor Andrzej Klawiter nigdy mi nie szczędził oraz za stworzenie przestrzeni, w której nawet najbardziej kontrowersyjne i niedopracowane pomysły mogły być swobodnie wyrażone i przedyskutowane.

Pracownikom Wydziału Psychologii i Kognitywistki, w szczególności uczestnikom seminarium instytutowego, którzy zawsze z dużą życzliwością dzielili się ze mną swoją wiedzą i uwagami dotyczącymi prezentowanych przeze mnie zagadnień.

Profesorowi Markowi Kaźmierczakowi, przyjacielowi, który zapraszając mnie na prowadzone przez siebie kilka lat temu seminaria dla studentów MISH - „przywrócił" mnie akademii. Udział w seminariach oraz zachęty Marka, by zamiłowanie do nauki przekształcić w „coś" wymiernego, były dla mnie impulsem, by "reaktywować" niedokończone podówczas studia filozoficzne i nawiązać współpracę z prof. Andrzejem Klawiterem.

Wreszcie, szczere wyrazy wdzięczności kieruję pod adresem Iwony, mojej Żony, która nie tylko dzielnie znosiła wyrzeczenia związane z pisaniem pracy, ale również dzieliła ze mną pasję do kognitywistyki i często oryginalnymi spostrzeżeniami twórczo wspierała mnie w konstruowaniu modelu działań intencjonalnych.

Spis treści

Streszczenie *9*

Summary *12*

1 Wprowadzenie i cel pracy *15*

1.1 Działania intencjonalne 15
1.2 Problem 26
1.3 O potrzebie multidyscyplinarnego podejścia do badania działań intencjonalnych 33
1.4 Teza 35
1.5 Cel i metoda 36
1.6 Zintegrowany model działań intencjonalnych (ZMDI) 37
1.7 Plan pracy 47

2 Przyczynowy wpływ stanów intencjonalnych na wybór zachowania *50*

2.1 Struktura intencjonalności 54
2.2 Schemat pełnego działania intencjonalnego 64
2.3 Proste i złożone działania intencjonalne 73
2.4 Działania podstawowe 76
2.5 Przyczynowy status intencji 78

3 Stany intencjonalne (idee) jako nagrody *94*

3.1 Hipoteza dopaminergicznego błędu predykcji nagrody (HDBPN) 97
3.2 Algorytm RL jako model obliczeniowy HDBPN 100
3.3 Algorytm RL jako podstawa złożonych działań intencjonalnych 111
3.4 Wybrane rozszerzenia metody uczenia się ze wzmacnianiem 130

4 Korelacyjno-interpretacyjny status stanów intencjonalnych towarzyszących prostym działaniom intencjonalnym *139*

4.1 Intencja w działaniu w ujęciu psychologii intencji 142
4.2 Poczucie sprawstwa 157
4.3 Funkcjonalne aspekty intencji oraz poczucia sprawstwa 175

5 Zintegrowany model złożonego działania intencjonalnego *188*

5.1 Problem naturalizacji umysłowych składników działania intencjonalnego 188
5.2 Cechy złożonego działania intencjonalnego 209
5.3 Zintegrowany model złożonego działania intencjonalnego 221

6 Zakończenie *302*

7 Bibliografia 307

8 Lista diagramów 328

9 Lista rysunków 329

10 Lista ilustracji 329

11 Legenda symboli 330

Table of contents

Summary (Polish version) *9*

Summary (English version) *12*

1 Introduction and purpose of the dissertation *15*

1.1 Intentional actions 15
1.2 Problem definition 26
1.3 The need for multidisciplinary studies of intentional actions 33
1.4 Thesis 35
1.5 Purpose and method 36
1.6 The integrated model of intentional actions (ZMDI) 37
1.7 Plan of the dissertation 47

2 The causal influence of intentional states on behavior *50*

2.1 The Structure of Intentionality 54
2.2 A diagram of a complete intentional action 64
2.3 Simple and complex intentional actions 73
2.4 Basic actions 76
2.5 The causal status of intention 78

3 Intentional states (ideas) as rewards *94*

3.1 The dopaminergic reward prediction error hypothesis (HDBPN) 97
3.2 Reinforcement Learning algorithm as a computational model of HDBPN 100
3.3 The RL algorithm as a basis for complex intentional actions 111
3.4 Selected extensions of the reinforcement learning algorithm 130

4 Correlational-interpretive status of intentional states accompanying simple intentional actions *139*

4.1 Intention to act in terms of the psychology of intention 142
4.2 Sense of agency 157
4.3 Functional aspects of intention and sense of agency 175

5 Integrated model of complex intentional action *188*

5.1 The problem of naturalizing the mental components of intentional action 188
5.2 Features of complex intentional action 209
5.3 Models 221

6 Conclusion *302*

7 References *307*

8 Diagrams 327

9 Figures 329

10 Pictures 329

11 Legend of Symbols 330

Streszczenie

STRESZCZENIE PRACY DOKTORSKIEJ

Zintegrowany model złożonych działań intencjonalnych

W rozprawie doktorskiej podejmuję problematykę modelowania działań intencjonalnych.

Celem rozprawy jest skonstruowanie zintegrowanego modelu działań intencjonalnych (ZMDI). Zadanie to wymagało zidentyfikowania najważniejszych mechanizmów i reprezentacji decydujących o przebiegu działań intencjonalnych oraz ustalenia zasadniczych powiązań między nimi. W konstrukcji modelu wykorzystano następujące koncepcje teoretyczne oraz wyniki badań eksperymentalnych: (1) teorię intencjonalności Johna Searle'a, (2) dane eksperymentalne zgromadzone w ramach psychologii intencji oraz (3) obliczeniowy model uczenia się ze wzmacnianiem opracowany przez neurobiologów na podstawie dopaminergicznego błędu predykcji nagrody.

Zaproponowane w dysertacji podejście odsłania zależności między poszczególnymi składowymi działania intencjonalnego, a także mechanizmy kontroli zachowań oraz kształtujące je procesy poznawcze.

Rozprawa doktorska składa się z pięciu rozdziałów.

W pierwszym rozdziale pokazano, jak rozproszony, niejednorodny i wycinkowy charakter mają teoretyczne oraz eksperymentalne badania działań intencjonalnych. Stosunkowo nieliczne próby systematyzacji zgromadzonego materiału ograniczone są do modeli jednodziedzinowych. W pracy przedstawiono argumenty uzasadniające potrzebę stworzenia modelu integrującego wiedzę o działaniach intencjonalnych, którą zgromadzoną na podstawie wyników badań przeprowadzonych przez uczonych reprezentujących różne dziedziny badawcze. Przedmiotem namysłu w dysertacji są złożone działania intencjonalne pojmowane jako system dwóch współdziałających mechanizmów:

(i) uczenia się ze wzmacnianiem oraz (ii) planowania na podstawie wiedzy zgromadzonej w formie sieci stanów intencjonalnych.

Wpływ stanów intencjonalnych na dobór zachowań został omówiony w **drugim rozdziale** pracy. W tej części rozprawy wykorzystano interpretację intencjonalności zaproponowaną przez Johna Searle'a. Czerpiąc inspirację z opracowanej przez amerykańskiego filozofa teorii intencjonalności, doprecyzowano i włączono do zintegrowanego modelu działań intencjonalnych nie tylko samą kategorię intencjonalności, ale również kluczowe typy stanów intencjonalnych (prior intencja, intencja w działaniu oraz przekonania). Efektem tych dociekań jest schemat działania intencjonalnego omawiany i komentowany w następnych rozdziałach dysertacji.

Rozdział trzeci poświęcony jest neurobiologicznemu mechanizmowi odpowiedzialnemu za organizowanie zachowań w uporządkowane sekwencje. Rdzeniem tego typu mechanizmu, zgodnie z hipotezą dopaminergicznego błędu predykcji nagrody (HDBPN), jest algorytm TDRL (Temporal Difference Reinforcement Learning), implementujący metodę uczenia się ze wzmacnianiem. W rozdziale tym omówiono zasadę działania algorytmu TDRL oraz jego najważniejsze cechy. Analizy modelu obliczeniowego uwzględniającego HDBPN prowadzą do określenia zakresu jego stosowalności w odniesieniu do złożonych działań intencjonalnych. Punktem wyjścia tych analiz jest hipoteza „super-mocy" Reada Montague'a, za pomocą której amerykański neuronaukowiec wyjaśnia specyficzne dla gatunku ludzkiego zachowania naruszające powszechnie przyjmowaną zasadę dominacji instynktu przetrwania.

W rozdziale czwartym poddano krytycznej analizie najważniejsze wyniki badań eksperymentalnych zgromadzonych przez psychologów intencji. Przywołane dane pozwalają wniknąć w strukturę zidentyfikowanej przez Searle'a intencji w działaniu oraz w jej „fenomenalne otoczenie". Z perspektywy celu niniejszej pracy najważniejszą częścią tego rozdziału jest analiza funkcjonalnych aspektów takich stanów jak: (1) poczucie chęci działania (sens of urge), (2) odniesienie do docelowego obiektu lub zdarzenia (reference forward to the goal object or event) oraz (3) poczucie sprawstwa (sens of agency).

W rozdziale piątym przedstawiono i szczegółowo objaśniono zintegrowany model złożonych działań intencjonalnych. Prezentacja modelu odbywa się w dwóch etapach. Etap pierwszy poświęcony jest sformułowaniu najważniejszych wymagań funkcjonalnych wobec modelu. Ich podstawą są wyniki wcześniejszych ustaleń. Etap drugi polega na

przedstawieniu zintegrowanego modelu w trybie kolejnych przybliżeń. Najpierw scharakteryzowany zostaje najprostszy model, w którym do kontroli zachowań wykorzystuje się jedynie mechanizm uczenia się ze wzmacnianiem. Następnie omawiane są modele coraz bardziej zaawansowane, w których uwzględnia się coraz więcej cech realnego działania intencjonalnego. Ostateczna wersja modelu pokazuje, jak wysoce złożone są struktury i mechanizmy działań intencjonalnych oraz – jak bardzo obraz ten odbiega od ich standardowych, monodziedzinowych ujęć.

Summary

THESIS SUMMARY

Integrated model of complex intentional actions

The dissertation addresses the issue of modeling intentional actions.

The purpose of this dissertation is to construct an integrated model of intentional actions (IMIA). It required the identification of the most important mechanisms and representations determining the course of intentional actions and the establishment of essential links between them. The following theoretical concepts and experimental results were used in the construction of the model: (1) John Searle's theory of intentionality, (2) experimental data collected within the psychology of intention, and (3) a computational model of reinforcement learning developed by neuroscientists in relation to the dopaminergic reward prediction error hypothesis. The proposed approach reveals the relationships among the various components of intentional action, as well as the mechanisms of behavioral control and the cognitive processes that shape them.

The dissertation consists of five chapters.

The first chapter shows how scattered, heterogeneous, and fragmented the theoretical and experimental studies of intentional actions are. While there have been relatively few attempts to systematize the collected material, they are limited to single-domain models. The study presents arguments justifying the need for a model that integrates knowledge about intentional actions obtained in different research fields. The subject of this chapter is complex intentional actions conceived as a system of two interacting mechanisms: (a) reinforcement learning and (ii) planning based on knowledge accumulated in the form of networks of intentional states.

Determining the influence of intentional states on behavior choice is the subject of **the second chapter** of the dissertation. It is based on John Searle's interpretation of intentionality. Drawing inspiration from the American philosopher's theory, the category of intentionality and the key types of intentional states (prior intentions, intentions in action and beliefs) have been defined and incorporated into an integrated model of intentional action. The result of these considerations is a conceptual framework of intentional action which serves as a reference in subsequent chapters of the dissertation.

The third chapter refers to the neurobiological mechanism responsible for organizing behavior into ordered sequences. The core of this type of mechanism, according to the dopaminergic reward prediction error hypothesis (HDBPN), is the Temporal Difference Reinforcement Learning (TDRL) algorithm, implementing the reinforcement learning method. In this chapter, I discuss the principle of operation and key features of the TDRL algorithm. Analyses of the computational model considering HDBPN lead to the determination of the scope of its applicability to complex intentional actions. The starting point for these analyses is the "superpower" hypothesis of Read Montague, by means of which the American neuroscientist explains human behavior that violates the commonly accepted principle of dominance of the survival instinct.

The fourth chapter consists analyzes of the most important experimental results gathered by psychologists of intention. The data brought to light provide insight into the structure of Searle's so-called “intention in action” and its "phenomenal milieu". The most important part of this chapter is the analysis of the functional aspects of such states as (1) sense of urge, (2) reference forward to the goal object or event and (3) sense of agency.

In **the fifth chapter**, I present and discuss in detail an integrated model of complex intentional actions. The model is presented in two stages. The first stage is devoted to formulating the most important functional requirements for this model. The second stage presents the integrated model of intentional action using an iterative approach. The simplest model, using only the reinforcement learning mechanism to control behavior, is characterized first. Subsequently, increasingly sophisticated models are discussed, which make it possible to incorporate additional features of real-life intentional action. The final version of the model shows how complex the structures and mechanisms of intentional actions are, and the extent to which this picture differs from the standard mono-domain account of these actions.

1 Wprowadzenie i cel pracy

W niniejszej rozprawie przedstawiam argumentację na rzecz tezy, że osiągnięcie postępu w badaniach nad działaniami intencjonalnymi wymaga odejścia od dotychczasowego, zgrubnego ich pojmowania i zastąpienia tradycyjnego, nadmiernie uproszczonego, schematu takiego działania przez model uwzględniający rozbudowaną charakterystykę jego wewnętrznej struktury. Twierdzę, że dopiero zintegrowany model działania intencjonalnego (ZMDI) odsłania jego zasadnicze składniki oraz subtelne zależności między nimi, których nie da się dostrzec poprzestając na dotychczasowych podejściach. Co więcej, proponowany w pracy model zachowuje niektóre z intuicji łączonych z intencjonalnością działania, nadaje im jednak precyzyjniejszą formę. Cel pracy, jakim jest skonstruowanie zintegrowanego modelu działania intencjonalnego, realizowany jest w kilku etapach. Etapy te przybliżają struktury zasadniczych składników działania intencjonalnego. W niniejszym rozdziale przedstawię standardową charakterystykę działania intencjonalnego, zasygnalizuję kłopoty teoretyczne, do których prowadzi takie jego pojmowanie oraz zaproponuję podejście alternatywne, wymagające stworzenia ZMDI. Omówię szkicowo cały model oraz jego najważniejsze składniki. Kolejne rozdziały są rozwinięciem ustaleń zasygnalizowanych w tym pierwszym.

1.1 Działania intencjonalne

W pierwszym, wstępnym przybliżeniu przez działanie intencjonalne rozumie się zachowanie[1], które wywołane jest przez zamiar podmiotu dążącego do osiągnięcia

[1] Za Bogdanem Sadowskim przyjmuję: „Pod pojęciem «zachowanie» rozumie się skoordynowane reakcje osobnika służące zaspokojeniu określonej potrzeby biologicznej, psychicznej lub społecznej zachodzące pod wpływem czynników wewnętrznych lub bodźców zewnętrznych. Formami zachowania mogą być zarówno proste reakcje ruchowe, jak kinezy, taksje i tropizmy, jak też u zwierząt wyższych złożone akty ruchowe, nabyte lub dziedziczne, nazywane reakcjami lub czynnościami behawioralnymi. Akty ruchowe mogą polegać na lokomocji (przemieszczaniu się) lub na manipulowaniu przedmiotami. Do czynności ruchowych należy też mimika i fonacja lub wokalizacja oraz pozy, na przykład grożenia lub uległości. […]

określonego celu. Jest to zachowanie jakiegoś osobnika, którego nie da się zrozumieć bez uwzględnienia jego „wewnętrznej" przyczyny, na którą składają się zamiar, czyli intencja jego podjęcia oraz zawarty w jej treści cel, czyli założony z góry skutek, do którego zachowanie to ma doprowadzić. Oprócz charakterystycznej struktury „intencja-przyczyna → zachowanie-skutek", działania intencjonalne cechują się ponadto dużym stopniem niezależności od napływających z otoczenia bodźców (w przeciwieństwie do wrodzonych odruchów), silnie zależą od kontekstu, w którym są realizowane, w szczególności od wcześniej wyuczonych asocjacji. Ponadto, realizację tego typu działań poprzedzają na ogół procesy planowania i rozumowania, a ich przygotowanie i wykonanie wymaga na ogół skupienia uwagi i wysiłku. Wreszcie, rezultaty działań intencjonalnych są ściśle monitorowane i stanowią podstawę do dalszego uczenia się (Haggard, 2005). Wymienione cechy powodują, że charakteryzują się one dużą elastycznością oraz różnorodnością zależną od możliwości poznawczych sprawcy. Przebieg i struktura tego typu działania przez wieki przykuwały uwagę filozofów, teologów oraz prawników. Z czasem, do badań włączyli się psychologowie i fizjologowie, a współcześnie dołączyli do nich neuronaukowcy i kognitywiści.

Introspekcja była przez wiele stuleci jedynym, uznawanym za prawomocny, sposobem wglądu w strukturę „wewnętrznej przyczyny" działań intencjonalnych. Niestety, zarówno sposób pojmowania introspekcji, jak i próby jej „stosowania" pozwalały jedynie na tworzenie nieostrych szkiców, co w efekcie uniemożliwiało wyodrębnienie i dokładny opis poszczególnych składników działania intencjonalnego. Tak używana introspekcja pozwalała co najwyżej stwierdzić, że działanie takie jest realizowane według prostego schematu: we wnętrzu, czyli umyśle danego osobnika pojawia się pewien zamiar (intencja), który, o ile osobnik nie ma zewnętrznych ograniczeń (np. kajdanek), przekształcany jest w działanie. W tym kontekście, źródłem zamiaru jest wola podmiotu traktowana jako byt wyjaśniający pierwszego rzędu, czyli taki, który służy wyjaśnieniu powiązanych z nim zjawisk, sam jednak nie podlega już wyjaśnieniu (Wegner, 2002, s. 12). Intencja pełni w tym procesie rolę przyczyny, a więc jest czynnikiem sprawczym, natomiast efekt, czyli zachowanie w postaci np. ruchu wraz z wywołanymi przez nie zmianami w świecie, pełnią rolę skutku. Innymi słowy, introspekcja upewnia podmiot o istnieniu przyczynowości mentalnej, której funkcjonowanie od strony filozoficznej w XVII wieku jako jeden z

Zachowanie służy ochronie przed niebezpieczeństwem, ułatwia poznawanie otoczenia, umożliwia rozród, opiekę nad potomstwem i tworzenie grup społecznych" (Sadowski, 2005, s. 21).

pierwszych w nowożytności, opisał René Descartes. Zgodnie z terminologią Kartezjusza – to, co umysłowe (*res cogitans*) może oddziaływać na to, co materialne (*res extensa*) (Descartes, 1958, s. 95–96), powodując określone zachowania[2]. Ten dualistyczny pogląd, obecny w różnych tradycjach religijnych i filozoficznych, wzmocniony koncepcjami zakładającymi istnienie nieśmiertelnej duszy oraz woli jako bytu pierwszego rzędu (tj. takiego, który wyjaśnia wszystko, nic natomiast go nie wyjaśnia), doprowadził w pewnym momencie, zdaniem Daniela Wegnera, do braku postępów w badaniu działań intencjonalnych (Wegner, 2002, s. 12).

Obraz aktów wolicjonalnych skomplikował się na przełomie XIX i XX wieku. Impulsem prowadzącym do zmiany myślenia na temat kontroli zachowań stały się powiązane z tym zagadnieniem rozważania dotyczące podmiotowości oraz wolności jednostki. Szczególnie wpływowe okazały się trzy nurty myślowe: (1) filozofia Karola Marksa, (2) filozofia Fryderyka Nietzschego oraz (3) psychoanaliza Zygmunta Freuda, razem klasyfikowane jako tzw. filozofie podejrzeń. Łączy je, pomimo zasadniczych różnic, podobny sposób myślenia o rzeczywistych przyczynach ludzkich myśli i zachowań. Wszyscy wymienieni myśliciele lokują owe przyczyny poza szeroko pojętą świadomością jednostki. Marks podkreślał w tym kontekście wpływ czynników społecznych, związanych z historią ludzkich form organizacji oraz form produkcji i wymiany. Nietzsche wskazywał na wolę mocy, a Freud na rolę nieświadomości oraz wpływ czynnika biologicznego, jakim był popęd seksualny (Felski, 2011). Można zatem stwierdzić, mówiąc metaforycznie, że w perspektywie filozofii podejrzeń „człowiek, […] nie działa, lecz jest działany przez anonimowe i wszechmocne siły" (Herling-Grudziński, 1990). Ukształtowana przez wymienione nurty filozoficzno-psychologiczne atmosfera intelektualna silnie wpłynęła na rozwój badań empirycznych dotyczących zachowań człowieka. Szczególne „piętno" na prowadzonych w tym obszarze dociekaniach odcisnęła metoda psychoanalityczna, która w istotny sposób zmieniła sposób myślenia o psychice ludzkiej. W szczególności, zmianie uległo myślenie o naturze relacji pomiędzy procesami świadomymi a nieświadomymi. Jak zauważa Erich Fromm: „zrozumienie własnej nieświadomości i niemożność pogodzenia

[2] „Uczy mnie także natura przez owe wrażenia bólu, głodu, pragnienia itd., że ja nie jestem tylko obecny w moim ciele, tak jak żeglarz na okręcie, lecz że jestem z nim najściślej złączony i jak gdyby zmieszany, tak że tworzę z nim jakby jedną całość." (Descartes, 1958, s. 95–96).

„Mówi się, że Kartezjusz postawił przed nami problem, w jaki sposób ruch w materialnym świecie może być nasycony lub ukształtowany przez umysł tak, aby mógł być uznany za działanie intencjonalne (*Descartes is said to have given us the problem of how a movement in the material world can be mind-imbued or mind-informed enough to count as an intentional act*)." (Baier, 1976, s. 27).

jej ze świadomym obrazem siebie samego - jest właśnie tym odkryciem, które nadaje psychoanalizie znaczenie radykalnego przedsięwzięcia zmierzającego w kierunku nowych form odkrywania siebie i nowej formy szczerości" (Fromm, 2006, s. 38). Oryginalny wkład psychoanalizy do badań nad psychiką ludzką nie uchronił tej metody, czy wręcz paradygmatu przed poważnymi błędami metodologicznymi. W rezultacie, oferowane przez psychoanalizę wyjaśnienia poddano ostrej krytyce metodologicznej (Popper, 1963). Podano także w wątpliwość rzetelność dostarczanych przez nią danych empirycznych oraz skuteczność propagowanych na jej gruncie metod terapeutycznych (Rakowska, 2005). Trudno byłoby wskazać wyniki badań przeprowadzonych przez psychoanalityków, oprócz przedstawionego przez Fromma ogólnikowego stwierdzenia dotyczącego relacji pomiędzy procesami świadomymi i nieświadomymi, które precyzowałyby wybrane mechanizmy odpowiedzialne za przebieg działań intencjonalnych.

Kontrowersje wokół rzetelności metody psychoanalitycznej przyczyniły się do ukształtowania behawioryzmu, nurtu, który zachowanie uczynił głównym przedmiotem badań psychologii eksperymentalnej, a stany umysłu zaklasyfikował jako nieistotne z perspektywy eksplanacyjnej – „psychologia nie jest nauką o umyśle" (Graham, 2017). Nietrudno zauważyć, że tak restrykcyjne podejście praktycznie uniemożliwiało pojmowanie, a w konsekwencji także i badanie, zachowań jako zależnych od stanów umysłowych (pragnień, intencji, itp.). Należy jednak podkreślić, że behawioryzm – pomimo tej „ślepoty" na procesy umysłowe – wniósł istotny wkład do nauki. W szczególności, stworzył teorię wzmocnień (*reinforcement theory*), zaproponowaną przez Skinnera, która przyczyniła się do rozpoznania związków, jakie zachodzą pomiędzy bodźcami warunkowymi, bezwarunkowymi a zachowaniem (patrz: schemat warunkowania klasycznego, warunkowania sprawczego, efekt generalizacji i różnicowania) (Zimbardo i in., 2010, s. 136). Obecnie, kiedy dysponujemy algorytmami modelującymi mechanizm uczenia się ze wzmacnianiem (Schultz i in., 1997; Sutton, 1998), to zidentyfikowane przez Pawłowa i Skinnera związki pomiędzy bodźcami a zachowaniami można wyjaśnić za pomocą precyzyjnie zdefiniowanych pojęć, takich jak: błąd predykcji nagrody, funkcja wartości czy współczynnik dyskonta (piszę o tym obszerniej w rozdziale 3. zatytułowanym „Stany intencjonalne (idee) jako nagrody").

Z czasem behawioryzm poddany został surowej krytyce. Wielu uczonych uznało, że postulowane w tym nurcie badawczym kryterium naukowości jest zbyt restrykcyjne i utrudnia czy wręcz uniemożliwia wyjaśnienie wielu zachowań. Dzieje się tak dlatego, że z

perspektywy behawioryzmu wyjaśnienie tzw. zachowań inteligentnych nie wymaga włączenia do eksplanansu stanów umysłowych podmiotów tych zachowań. Ponieważ działania intencjonalne są podzbiorem zachowań inteligentnych, dlatego w ich przypadku nie jest potrzebne odwoływanie się do stanów umysłowych, które – zdaniem zwolenników takiego podejścia – mają albo status bytów fikcyjnych (takich np. jak flogiston przed odkryciem tlenu), albo mogą być sprowadzone do stanów organizmu (włączając do nich procesy mózgowe), wywołanych przez określone sytuacje zachodzące w otoczeniu. Z perspektywy współczesnej postawa taka wydaje się zasadna. Wszak dzisiaj, kiedy dysponujemy już aparaturą do nieinwazyjnego rejestrowania aktywności w systemach mózgowych, badacze podejmują próby wyjaśniania wybranych działań intencjonalnych przez wskazanie poprzedzających je bądź towarzyszących im procesów w mózgu. Narzędziami takimi nie dysponowali jednak twórcy behawioryzmu, a także ich następcy, dlatego ich program badawczy w dużej mierze miał charakter deklaratywny. Co więcej, nawet dzisiaj, gdy dysponujemy skanerami mózgu, trudno spotkać poważnych badaczy, którzy broniliby generalnej tezy, że wyjaśnienie działania intencjonalnego nie wymaga odwołania się do stanów umysłowych jego podmiotu. W połowie lat 50. XX wieku badacze z różnych dyscyplin uświadomili sobie, że nie warto czekać na hipotetyczną sytuację z przyszłości, kiedy to okaże się, że postępy w badaniu mózgu pozwolą zastąpić wiedzę o procesach umysłowych zaawansowaną wiedzą o jego stanach. Zamiast tego, podjęli wysiłki, aby – korzystając z tego, co już udało się wypracować – stworzyć zintegrowaną naukę o umyśle, która uniknie „mielizn" behawioryzmu. Kluczową rolę w tym przedsięwzięciu odegrali: George Miller – psycholog, John McCarthy, Marvin Minsky, Allen Newell i Herbert Simon – badacze z obszaru sztucznej inteligencji i procesów obliczeniowych oraz Noam Chomsky – twórca nowatorskiej teorii językoznawczej. Badacze ci, a także wielu innych, zakwestionowali behawiorystyczne założenie o nieistotności stanów umysłowych (Thagard, 2014), przywracając należny im w dociekaniach psychologicznych status, a zarazem zapoczątkowali nowy paradygmat w badaniach nad umysłem, który polega na tworzeniu obliczeniowych modeli jego działania. W ten sposób narodziła się kognitywistyka.

Rekonstrukcja stojących za poszczególnymi zachowaniami procesów mózgowych oraz powiązanych z nimi reprezentacji znacząco przyspieszyła, gdy badacze zaczęli korzystać z coraz bardziej zaawansowanych metod eksperymentalnych oraz wspierających je narzędzi. Istotnym zdarzeniem w historii badań nad działaniami dowolnymi było wykorzystanie

encefalografu (EEG) oraz elektromiografu (EMG). Przełomowe okazały się dwa eksperymenty: (1) odkrycie za pomocą EEG czasowego przebiegu potencjału gotowości dla spontanicznego zgięcia nadgarstka (Kornhuber & Deecke, 1965) oraz (2) odkrycie, że potencjał gotowości do wykonania ruchu poprzedza moment, w którym pojawia się świadoma chęć do jego zainicjowania (Libet, Gleason, Wright, & Pearl, 1983). Te ostatnie wyniki zainspirowały wielu badaczy do nowych prac nad działaniami intencjonalnymi. W ten sposób wykształciła się m.in. psychologia intencji, która stopniowo zaczęła wnikać w złożoną naturę prostych z pozoru zachowań. Z czasem, poza EEG i EMG oraz tradycyjnymi narzędziami psychologii eksperymentalnej, subdyscyplina ta zaczęła wykorzystywać nowe narzędzia do nieinwazyjnego badania aktywności mózgu, takie jak: bezpośrednia, przezczaszkowa stymulacja mózgu, (TMS, *transcranial magnetic stimulation*), magnetoencefalografia (MEG, *magnetoencephalography*) czy obrazowanie funkcjonalne za pomocą rezonansu magnetycznego (fMRI, *functional magnetic resonance imaging*). Stosowano je zarówno do badania typowych, niezakłóconych procesów poznawczych, jak i do badania różnego rodzaju zakłóceń powstałych w wyniku uszkodzenia określonych struktur mózgu. Z każdym rokiem pojawiają się nowe dane, które dają asumpt do poszukiwania całościowego modelu działań intencjonalnych.

W systematycznych analizach procesów podejmowania działań przeprowadzanych m.in. w psychologii, badacze starają się odejść od potocznych, a także czysto filozoficznych, oderwanych od aktualnej wiedzy naukowej sposobów pojmowania intencji. We współczesnej nauce rozumie się intencję tak, jak sformułował to Patryk Haggard: „Termin «intencja» obejmuje szereg różnych, powiązanych ze sobą procesów, w których przetwarzanie informacji prowadzi do przekształcenia pragnień i celów w zachowanie.”[3]. Zaproponowane przez Haggarda określenie dobrze oddaje złożoną naturę działań intencjonalnych. Wskazuje ono wyraźnie, że ich realizacja wymaga całego szeregu procesów przetwarzania informacji zorganizowanych tak, aby zamierzony cel (pożądany stan rzeczy), będący momentem początkowym całego łańcucha zjawisk, przekształcony został w zachowanie, mające doprowadzić do jego urzeczywistnienia. Widać zatem, że tego typu działania można traktować jako splot, w którym punktami węzłowymi są: (1) stan intencjonalny (np. pragnienie, ujęte łącznie z przedmiotem, ku któremu jest skierowane) oraz (2) zachowanie podejmowane ze względu na przedmiot intencjonalny

[3] “The term ‘intention’ covers several distinct processes within the chain of information processing that translates desires and goals into behavior.” (Haggard, 2005, s. 290).

(np. sekwencja celowych ruchów). Ujęcie Haggarda jest jednak tylko pierwszym przybliżeniem specyfiki działania intencjonalnego. Powyższa charakterystyka będzie, zdaniem tego badacza, niepełna, jeśli nie wniknie się w strukturę intencji oraz w procesy stanowiące jej najbliższe otoczenie, czyli w określone przeżycia i konteksty, które wpływają na nasze zachowania. We współczesnych badaniach nad działaniami intencjonalnymi wyróżnia się dwa stany, które mają związek z realizacją zachowań podlegających świadomej kontroli: (1) **prior intencję** oraz (2) **intencję w działaniu.** Ta ostatnia ma dwie składowe: (a) *poczucie chęci* [wykonania ruchu] (*sense of urge*) oraz (b) *skierowanie ku docelowemu obiektowi lub zdarzeniu* (*reference forward to the goal object or event*) (Haggard, 2005). Kiedy w literaturze przedmiotu opisującej działania intencjonalne pisze się o: „przeżyciu intencjonalności działania" (*experience of intentionality*), „poczuciu wysiłku" (*experience of effort*), „przeżyciu przyczynowości mentalnej" (*experience of mental causation*), „poczuciu sprawstwa" (*sense of agency*) czy o powiązanym z takimi przeżyciami doświadczeniu wolnej woli (*experience of free will*), to ma się na uwadze albo kombinację wszystkich tych składników, albo tylko niektórych z nich (Bayne, 2006, s. 170). Wskazane stany intencjonalne „dodają" do zachowania podmiotu element jakościowy, który sprawia, że nie jest już ono postrzegane wyłącznie jako sekwencja mechanicznych ruchów, lecz traktowane jest jako „nośnik" informacji o przyczynie mentalnej (stanie lub sekwencji stanów intencjonalnych), która je wywołała. „Odczytanie" tej informacji przez obserwatora pozwala mu na włączenie takiego zachowania w określone konteksty poznawcze, emocjonalne i społeczne.

Intrygujące jest to, że stosunkowo mało uwagi poświęca się dokładniejszej charakterystyce tych stanów intencjonalnych. W najlepszym razie mamy do czynienia z luźnymi uwagami, jakie nasuwają się badaczom, kiedy przystępują do interpretowania wyników uzyskanych w eksperymentach. Tymczasem potrzeba teorii, która w spójny sposób wyjaśniłaby zależności istniejące między stanami intencjonalnymi oraz określiła ich wpływ na podjęcie i przebieg działania intencjonalnego. W szczególności, gruntownego przemyślenia wymaga status sygnalizowanego wcześniej, introspekcyjnie dostępnego przekonania o ich przyczynowej roli. Rozważanie takie powinno uwzględnić zarówno zgromadzone dane empiryczne, jak i dostępne koncepcje teoretyczne. Propozycję całościowego ujęcia wskazanej kwestii przedstawił w 2002 roku psycholog z Uniwersytetu Harvarda Daniel Wegner w książce zatytułowanej *The Illusion of conscious will* (Wegner, 2002). Z perspektywy głównego celu dysertacji, czyli konstrukcji zintegrowanego modelu

działań intencjonalnych, opracowanie amerykańskiego psychologa pomaga doprecyzować status wybranych stanów umysłowych (tzw. stanów fenomenalnych) poprzedzających działania lub towarzyszących im, oraz opisuje szereg mechanizmów zaangażowanych w świadome podejmowanie decyzji. Niektóre ze spostrzeżeń Wegnera zostaną wykorzystane w modelu zaproponowanym w 5 rozdziale rozprawy.

Wegner rozpoczyna rozważania od analizy danych zgromadzonych przez Libeta (Libet i in., 1983a) oraz proponuje własną ich interpretację:

> *Nie wiemy, jakie konkretne nieświadome procesy umysłowe może reprezentować RP (potencjał gotowości). […]. Miejsce świadomej woli na osi czasu zdaje się sugerować, że jest to doświadczenie będące ogniwem w łańcuchu przyczynowym prowadzącym do działania. W rzeczywistości jednak może nie być nawet tym. Świadoma wola może po prostu nie mieć nic do roboty. Jest ona, podobnie jak samo działanie, wywołana przez wcześniejsze zdarzenia mózgowe i umysłowe.*[4]

Wegner zaproponował, by uznać, że świadoma wola towarzysząca zachowaniu (w psychologii intencji zwana „poczuciem sprawstwa" (*sense of agency*)) to nadbudowana nad intencją w działaniu szczególnego rodzaju emocja, którą za Damasio zaklasyfikował jako tzw. marker somatyczny (Damasio, 2011). W opinii Wegnera, jej główną funkcją jest ułatwianie działającemu podmiotowi identyfikowania wykonanych przez niego działań (*authorship, sense of agency*). U podstaw tej propozycji leży następująca obserwacja: istnieją przypadki, gdy obserwowane z zewnątrz działania wyglądają na dowolne i intencjonalne, jednak z perspektywy ich wykonawców traktowane są jako niechciane i mimowolne (np. zespół obcej ręki czy zespół Tourette'a (Wegner, 2002, s. 4)). Zdarza się, że działania uznawane przez sprawcę za zamierzone i umyślne są tak naprawdę działaniami impotentnymi, nad którymi nie ma on rzeczywistej kontroli (np. nie można kontrolować joysticka w grze komputerowej działającej w trybie demo, choć można wywołać złudzenie, że kontrolę nad nim się sprawuje). Innymi słowy, poczucie sprawstwa jest niezależnym komponentem procesu organizującego działanie. W zależności od „natężenia" tego poczucia dane działanie może jawić się jako chciane i zamierzone albo niechciane czy wręcz zupełnie przypadkowe.

[4] "We don't know what specific unconscious mental processes the RP might represent.... The position of conscious will in the timeline suggests perhaps that the experience of will is a link in a causal chain leading to action, but in fact it might not even be that. It might just be a loose end — one of those things, like the action, that is caused by prior brain and mental events." (Wegner, 2002, s. 55).

Powyższe spostrzeżenie skłoniło Wegnera do skonstruowania modelu działania intencjonalnego składającego się z następujących procesów:

- Nieświadomego, mózgowego procesu przygotowawczego, którego efektem jest określony ruch ciała.
- Nieświadomego procesu odpowiedzialnego za wyznaczenie myśli związanej z działaniem.
- Procesu interpretowania polegającego na powiązaniu w kategoriach mentalnej przyczynowości obiektu, jakim jest myśl, wyznaczona w nieświadomym procesie (przyczyna) z przewidywanym lub zaobserwowanym ruchem ciała (skutek), będącym następstwem mózgowego procesu przygotowawczego. W typowym przypadku, efektem działania takiego procesu interpretowania jest szczególnego rodzaju, nacechowany emocjonalnie, stan umysłu – tzw. poczucie sprawstwa. W przypadkach nietypowych poczucie to zostaje całkowicie wytłumione, a w konsekwencji działanie nie jest uznane za chciane (patrz: zespół obcej ręki) lub jest przypisane innemu agentowi.

  Przeżycie świadomej woli powstaje wówczas, gdy wnioskujemy, że nasza świadoma intencja była przyczyną naszego dobrowolnego działania, pomimo, że zarówno intencja, jak i działanie wywołane zostały przez procesy umysłowe, których nie odczuwamy jako chciane.[5]

Wegner poddał powyższy model weryfikacji eksperymentalnej. W specjalnie zaprojektowanym eksperymencie *I Spy* (jego szczegółowy opis znajduje się w rozdziale 4.) wykazał, że proces interpretacji działania jest niezależny od jego faktycznych przyczyn. Niezależność, zdaniem Wegnera, polega na tym, że przyczyn zachowania poszukuje się w towarzyszących działaniu myślach, choć w rzeczywistości to nie one wywołują działanie. Innymi słowy, przekonanie o kontrolnej funkcji intencji to iluzja. W takim ujęciu, podmiot *post factum* interpretuje własne zachowania oraz ich konsekwencje (tj. zmiany w otoczeniu) jako wywołane przez stany umysłu, które się pojawiły na etapie przygotowania do działania lub w trakcie jego realizacji. Postępując w ten sposób ulega on iluzji, gdyż stany te same z siebie nie mają mocy sprawczej, a jedynie poprzedzają lub towarzyszą

[5] "The experience of conscious will arises when we infer that our conscious intention has caused our voluntary action, although both intention and action are themselves caused by mental processes that do not feel willed." (Wegner, 2002, s. 55).

działaniu, a więc to nie one są faktyczną przyczyną zachowania. Wegner przytacza w tym kontekście obrazową metaforę:

> *Czy kompas steruje statkiem? W pewnym sensie można powiedzieć, że tak, ponieważ pilot odwołuje się do kompasu, ustalając, czy należy wprowadzić zmiany w kursie statku. Jeśli wygląda na to, że statek płynie na zachód w kierunku skalistego brzegu, należy skręcić na północ do portu, by uniknąć nieszczęścia. Oczywiście, kompas w żadnym fizycznym sensie nie steruje statkiem. Igła po prostu ślizga się w obudowie kompasu, niczym nie sterując. W związku z tym kuszące wydaje się odesłanie małej magnetycznej igły do klasy epifenomenów – rzeczy, które nie mają wpływu na to, dokąd statek popłynie.*[6]

W ten sposób koncepcja Wegnera wpisuje się w szerszy kontekst teoretyczny, w tzw. interpretacjonizm, tj. pogląd redukujący stany intencjonalne do roli komentarzy lub wyjaśnień *ad hoc* powoływanych do życia w celu zaspokojenia określonych potrzeb psychicznych podmiotu, np. potrzeby komfortu poznawczego (Festinger, 1957) czy potrzeby wyjaśnienia i zrozumienia własnych zachowań, w szczególności w kontekście społecznym (Bem, 1967; Gazzaniga, 1978). Taki pogląd ma swoje źródła w wielu eksperymentach psychologicznych, w których wykazano, że faktyczne przyczyny zachowań są często niedostępne ich podmiotom. Badani, którym zadawano pytanie: „Co spowodowało, że podjęli oni takie, a nie inne działanie?", odpowiadali konstruując wyjaśnienia *ad hoc* niemające wiele wspólnego z faktycznymi przyczynami ich zachowania, stąd teza Wegnera o iluzyjnym charakterze tego typu wyjaśnień, a w konsekwencji również świadomej woli jako iluzji. W słynnym artykule *Telling More Than We Can Know: Verbal Reports on Mental Processes*, Richard Nisbett i Timothy Wilson (Nisbett & Wilson, 1977) dokonali przeglądu tego typu eksperymentów i wskazali na ograniczony zakres dostępu introspekcyjnego do czynników determinujących zachowanie. Wskazane przez Nisbetta i Wilsona przypadki znalazły swoje teoretyczne opracowanie w licznych pracach z obszaru psychologii. Teorie dysonansu poznawczego Festingera,

[6] "Does the compass steer the ship? In some sense, you could say that it does, because the pilot makes reference to the compass in determining whether adjustments should be made to the ship's course. If it looks as though the ship is headed west into the rocky shore, a calamity can be avoided with a turn north into the harbor. But, of course, the compass does not steer the ship in any physical sense. The needle is just gliding around in the compass housing, doing no actual steering at all. It is thus tempting to relegate the little magnetic pointer to the class of epiphenomena — things that don't really matter in determining where the ship will go." (Wegner, 2002, s. 317).

samoobserwacji Bema czy lewopółkulowego interpretatora Gazzanigi to przykłady tego typu konstrukcji.

Rozwiązanie Wegnera, w którym psycholog odwołuje się do iluzyjnej natury intencji i poczucia sprawstwa jest oryginalną próbą wyjaśnienia działań intencjonalnych. Powstaje jednak pytanie: czy zasadne jest rozciągnięcie zaproponowanego wyjaśnienia na wszystkie przypadki takich działań?

W szczególności, wątpliwe wydaje się stanowisko, że świadome zamiary oraz inne towarzyszące działaniom stany intencjonalne dostosowują się tylko do niepodlegających świadomej kontroli procesów organizujących działania intencjonalne. Konsekwencją takiego poglądu jest teza Wegnera mówiąca, że świadoma wola ma charakter iluzyjny i konstruktywistyczny. W mojej opinii, stanowisko takie jest wątpliwe, gdyż intencję redukuje się wyłącznie do aktu działania, a pomija się w nim inny jej ważny aspekt, tj. konstruowanie planu realizacji takiego działania. Choć plany i związane z nimi intencje nie zawsze prowadzą do realizacji działań (tzn. nie zawsze są skuteczne przyczynowo), to bez nich w wielu przypadkach niemożliwe byłoby zrozumienie ludzkich wyborów i zachowań. Gdy ktoś, kogo znamy, czeka na lotnisku na lot do Singapuru, często z góry jesteśmy w stanie określić powód, dla którego zdecydował się na tego typu podróż. Jeśli jest on pracownikiem globalnej korporacji, to prawdopodobnie jest w delegacji. Jeśli jest naukowcem, to przypuszczalnie bierze udział w konferencji naukowej lub został zaproszony, by wygłosić cykl wykładów. W innym przypadku przyczyną podróży może być wycieczka w region świata, który od zawsze go interesował. Wskazane powody uwzględniają nasze wcześniejsze doświadczenia oraz wiedzę odnoszące się do danej osoby i do świata. Przykład ten świadczy o przyczynowo-wyjaśniającej funkcji intencji. Z jednej strony pozwala uruchomić stosowną do potrzeb sekwencję działań, z drugiej strony jest ważnym źródłem informacji wyjaśniającym zaobserwowane u innych agentów[7] zachowania. By uzgodnić stanowisko Wegnera z ujęciem traktującym intencję jako element procesu planowania, w pracy wykorzystane zostanie wprowadzone powyżej rozróżnienie na: prior intencję oraz intencję w działaniu.

[7] W dysertacji przyjmuje się następującą definicję agenta: agent to istota zdolna do działania (*an agent is a being with the capacity to act* (Schlosser, 2019)). Obejmuje ona zarówno agentów sztucznych np. roboty, jak i agentów biologicznych.

1.2 Problem

Należy zauważyć, że koncepcji Wegnera towarzyszy pewna istotna trudność. Gdyby przyjąć, że przekonania, pragnienia, intencje, lęki, nadzieje, które uznajemy za przyczyny lub składniki przyczyn naszych działań, mają głównie charakter konstruktywistyczny i jedynie „dopowiadają" do zachowań pewną historię, to uzyskalibyśmy wątpliwy z punktu widzenia ewolucji naszego gatunku obraz podmiotu działania intencjonalnego. To, co od strony energetycznej jest bardzo kosztowne (Clarke & Sokoloff, 1999), czyli wytwarzanie świadomego stanu intencjonalnego każdorazowo, kiedy podejmowane jest działanie inteligentne - byłoby jedynie „ornamentem" działania, pozbawionym wpływu na jego przebieg. Tak pojmowana intencja nie pełniłaby funkcji przystosowawczej. Nieco przerysowując można by powiedzieć, że w interpretacjonizmie, a tak klasyfikuje się stanowisko Wegnera, podmiot działający traktuje się jako rodzaj „automatu" z wbudowaną opcją komentatora, pozbawiając tym samym świadome stany umysłowe mocy sprawczej. Zgodnie z tą koncepcją, komentarze odnoszące się do działań uczestniczą w tworzeniu reprezentacji umysłowych, jednak nie mają wpływu ani na ich inicjowanie, ani na ich kontrolę. Cechy organizmu nieposiadające funkcji przystosowawczej nie muszą być selekcjonowane w procesie ewolucji (patrz: koncepcja tzw. spandreli, wprowadzona do biologii ewolucyjnej przez Stephena Jaya Goulda (Gould & Lewontin, 1979)), jednak w dłuższej perspektywie czasowej, kosztowne od strony energetycznej „rozwiązania" ewolucyjne (w tym przypadku są nimi świadome stany intencjonalne) zostają wyeliminowane, jeśli nie zwiększają szans na przetrwanie (por. Searle, 1983, s. 160).

Konstruktywistyczne stanowisko Wegnera budzi wątpliwości nie tylko z perspektywy kosztów energetycznych ponoszonych przez organizm, ale również niezgodne jest z podstawową logiką funkcjonowania stanów intencjonalnych, zrekonstruowaną przez Johna Searle'a. Stany intencjonalne, zdaniem amerykańskiego filozofa, to nic innego, jak reprezentacje o specyficznym odniesieniu oraz strukturze. Chciałbym, nie wchodząc w tym miejscu w szczegóły teorii intencjonalności (omawiam ją dokładniej w rozdziale 2), odwołać się do wskazanej przez Searle'a podstawowej funkcji stanów intencjonalnych, tj. do ich zdolności odnoszenia do środowiska. Amerykański filozof zakłada, że sieć stanów intencjonalnych to szczególnego rodzaju baza wiedzy, za pomocą której agent orientuje się w świecie i – na podstawie której – próbuje organizować swoje zachowania. Jeśli uznać, iż jest to podstawowe zadanie stanów intencjonalnych, to przyjąć trzeba, że skłonność podmiotu do tego, by „kreować określone iluzje na temat siebie oraz środowiska", musi

być ściśle kontrolowana i limitowana. Trudno wyobrazić sobie, by organizm wykorzystujący stany intencjonalne wyłącznie do „kreowania" wyjaśnień na własny temat – mógł sprawnie funkcjonować w środowisku. Byłoby to niejako zaprzeczeniem podstawowej funkcji tych stanów. Stany intencjonalne, zamiast pomagać nam w odwzorowywaniu świata oraz występujących w nim związków przyczynowych lub wspierać nas w „dostosowywaniu" świata do naszych oczekiwań, tworzyłyby zbiór „spójnych" komentarzy w „arbitralny" sposób odnoszących się do rzeczywistości. Należy równocześnie zauważyć, że tak określona podstawowa funkcja stanów intencjonalnych nie wyklucza tego, że część z nich jest po prostu błędna. Zjawisko halucynacji w systemie poznawczym jest dobrym przykładem obrazującym problem adekwatności reprezentacji umysłowych. Searle twierdzi, wbrew opinii wielu filozofów umysłu, że tego typu przeżycie zdarza się zdrowym osobom niezwykle rzadko, w pewnym sensie jest wyjątkowe, a związana z nim błędna reprezentacja rzeczywistości zostaje szybko rozpoznana. Przeżycie takie jest na ogół traktowane jako pomyłka, przesłyszenie lub przewidzenie i świadczy ono o powszechnej adekwatności reprezentacji umysłowych (Searle, 2011). Gdy te statystycznie rzadkie sytuacje staną się normą, wówczas podlegający halucynacjom podmiot znajdzie się w niebezpiecznym położeniu. Wyraźnie obrazują to zachowania osób, u których zdiagnozowano poważne dysfunkcje systemu poznawczego, np. u chorych na schizofrenię. W takich przypadkach liczba reprezentacji nieadekwatnych jest na tyle duża, że działania podejmowane przez osoby cierpiące na tego typu chorobę są nieefektywne, a często zagrażają ich życiu.

Searle twierdzi również, w przeciwieństwie do eliminatywistów (np. Patricii i Paula Churchlandów (P. M. Churchland, 1998)), że stany umysłowe pełnią rolę przyczynową, innymi słowy przyczynowość intencjonalna jest możliwa (rozumiana jest ona przez Searle'a inaczej, niż w kartezjańskim interakcjonizmie (Searle, 1983, s. 118)). Określone stany intencjonalne, np. prior intencje oraz intencje w działaniu, realnie wpływają na wybór naszych zachowań – stany te są przyczynowo-skuteczne. Searle twierdzi, wbrew Hume'owi, że przyczynowość intencjonalna nie jest złudzeniem, wręcz przeciwnie – jest jednym z naszych podstawowych sposobów doświadczania świata. Na tym poziomie ogólności można stwierdzić, że ujęcie amerykańskiego filozofa jest niejako antytezą stanowiska Wegnera, będącego twórcą koncepcji pozornej przyczynowości umysłowej (*apparent mental causation*).

Zaprezentowane stanowiska – Wegnera i Searle'a – reprezentują dwa przeciwstawne nurty badawcze w dziedzinie działań intencjonalnych. Pierwszy, głównie empiryczny, traktuje stany intencjonalne towarzyszące zachowaniom jako składowe iluzji, czyli skonstruowanego przez proces interpretacyjny wyjaśnienia dotyczącego intencji zrealizowanego działania. Drugi, wywodzący się głównie z filozofii umysłu, uznaje przyczynową sprawczość (lub przynajmniej współsprawczość) stanów intencjonalnych, w szczególności takich stanów, jak prior intencja oraz intencja w działaniu. Obydwa, tak wyraźnie odmienne, stanowiska badawcze zawierają idee, które warte są przeanalizowania. W niniejszej rozprawie przedstawię propozycję uzgodnienia tych – z pozoru wykluczających się – koncepcji teoretycznych oraz włączę je w zaproponowany przeze mnie zintegrowany model działania intencjonalnego.

W tym kontekście pojawia się fundamentalne pytanie: jaki jest faktyczny status stanów intencjonalnych oraz przeżyć fenomenalnych towarzyszących zachowaniom składającym się na dane działanie intencjonalne? Uważam, że zaproponowane przez Wegnera ujęcie roli stanów umysłu jako komentarzy do zachowań jest problematyczne i wymaga korekty. Uznaje on bowiem, że postanowieniu wykonania działania D (Wegner określa ten akt umysłowy za pomocą potocznej nazwy „myśl" [*thought*]) i następującej po nim faktycznej realizacji działania [*action*] D towarzyszy doznanie [*experience*] świadomej woli. Doznanie to jest podstawą przekonania, że wystąpienie postanowienia – myśli – jest przyczyną fizycznego działania. W koncepcji Wegnera iluzją nie jest sama czynność umysłowa, którą jest postanowienie o podjęciu działania, np. decyzja, że teraz podniosę rękę. Iluzoryczne jest samo doświadczenie [*experience of conscious will*], sprowadzające się do nieodpartego wrażenia, a w efekcie także i przekonania, że to właśnie czynność umysłowa, mająca postać postanowienia, jest siłą sprawczą wywołującą fizyczny ruch mojej ręki. Nie kwestionuje on tego, że w naszych umysłach pojawiają się szczególnego typu myśli o podjęciu określonego działania. Wegner twierdzi natomiast, że podmiot rejestruje nie tylko dwa odrębne zdarzenia: myśl i cielesne działanie, ale także istniejący między nimi związek następstwa. Na podstawie tego związku wyspecjalizowany proces poznawczo-emocjonalny konstruuje pomiędzy myślą a działaniem relację, która objawia się (1) na poziomie treściowym: w formie przekonania, że przyczyną działania jest towarzysząca mu myśl (intencja), zaś (2) na poziomie fenomenalnym: w formie poczucia sprawstwa (*sense of agency*), czyli szczególnego rodzaju emocji tła. I właśnie treść przekonania jest – zdaniem Wegnera – iluzją. Ujęcie takie przypomina Hume'owską

analizę związku przyczynowego. Nie jest to przypadkowe, gdyż Wegner nawiązuje w swoich wywodach do argumentacji szkockiego filozofa.[8] Koncepcja Wegnera wymaga szerszego omówienia i krytycznej analizy. Zajmę się tym w rozdziale 4., teraz jedynie zaznaczę, że Wegnerowskie ujęcie związku między procesami umysłowymi a działaniami (zachowaniami) pomija dwie istotne kwestie: (1) złożoną strukturę zespołu procesów umysłowych, którą nazywa, w dużym uproszczeniu, myślą oraz (2) neuronalne podłoże tych procesów.

Przyjęta przez Wegnera strategia, by oprzeć model działania intencjonalnego na wyniku Libeta oraz zidentyfikowanym przez niego „opóźnieniu" intencji wykonania ruchu względem poprzedzającego ją potencjału gotowości motorycznej, jest dość ryzykowna. Wegner „rozciąga" wspomniany efekt, występujący w prostych działaniach intencjonalnych, na przypadki złożone. Założenie, że w przypadku złożonych działań intencjonalnych występuje taki sam mechanizm „opóźnienia" jest wątpliwe i jeśli Wegner je podziela, to powinien to wyraźnie stwierdzić i przedstawić stosowne uzasadnienie. Jest to niezbędne dlatego, że zdecydowana większość działań intencjonalnych – to działania złożone, a nie proste. Pod tym względem nauka o zachowaniach systemów intencjonalnych nie odbiega od innych nauk zajmujących się systemami złożonymi. Strategia, jaką się przyjmuje w takich przypadkach, nie polega na badaniu przypadków prostych, by następnie multiplikować układy proste po to, by tworzyć z nich złożone. Znacznie efektywniejsze jest opisanie przypadków prostszych jako tzw. przypadków brzegowych pewnego ogólnego mechanizmu realizującego dowolną możliwość.

Omawiane wyżej koncepcje (Searle'a i Wegnera), choć tak odmienne, są pod jednym, istotnym względem podobne. Ich autorzy, formułując odpowiedź na pytanie o naturę działań intencjonalnych, skupiają się na wysokopoziomowych strukturach i mechanizmach

[8] „Osoba doświadczająca woli jest, zgodnie z tym poglądem, w takiej samej pozycji, jak ktoś, kto spostrzega przyczynowość, obserwując uderzenie jednej kuli bilardowej w drugą. Nauczyliśmy się od Hume'a, że o przyczynowości w kręglach, bilardzie czy innych grach wnioskuje się na podstawie stałego połączenia ruchów kul. Dlatego zasadne jest przyjęcie, że o woli – doświadczeniu własnej przyczynowej sprawczości – wnioskuje się na podstawie połączenia zdarzeń prowadzących do działania. … Jakież obiekty w naszych umysłach zdają się nam zderzać tak, że wywołują spostrzeżenie woli? … Skłonni jesteśmy uznawać siebie za autorów działania przede wszystkim wtedy, kiedy wcześniej, w stosownym przedziale czasowym, doświadczyliśmy istotnych myśli o tym działaniu. Pozwala nam to wnioskować, że nasze własne procesy umysłowe wprawiły działanie w ruch. Podejmowane przez nas działania, które nie są zapowiedziane wcześniej w naszych umysłach, jawią się nam jako niewywołane przez nasz umysł. Nasze zamiary (intencje), aby działać, mogą, ale nie muszą *być* [faktycznymi – dodatek MC] przyczynami. Nie ma to jednak znaczenia, gdyż najważniejsze jest to, abyśmy *spostrzegali* je jako przyczyny, jeśli mamy doświadczyć świadomej woli." (Wegner, 2002, s. 64-65).

umysłowych, pomijając niskopoziomowe procesy, które leżą u podłoża zjawisk pojawiających się w umyśle. Przez podłoże rozumiem tu mechanizmy wbudowane w systemy neuronalne, których aktywność powiązana jest z pojawieniem się w umyśle świadomego zamiaru, a w konsekwencji – decyzji o podjęciu działania mającego doprowadzić do realizacji celu, który jest treścią zamiaru.

Podkreślić chciałbym, że moje wątpliwości co do trafności omówionych wyżej koncepcji działań intencjonalnych nie dotyczą tego, że badacze ci pomijają związek między procesami mózgowymi a tymi procesami umysłowymi, które zaangażowane są w podejmowanie działań intencjonalnych. Wielu referowanych tu autorów, np. John Searle, często *explicite* uznaje taki związek. Jednak zarówno Searle, jak i Wegner wyjaśniając działania intencjonalne skupiają swoją uwagę wyłącznie na ich wysokopoziomowych cechach, pomijając w swoich analizach modele obliczeniowe odnoszące się do niskopoziomowych procesów odpowiedzialnych za ich przebieg. Moim zdaniem, nieuwzględnienie tego typu modeli w odniesieniu do działania intencjonalnego uniemożliwia efektywne jego wyjaśnienie. Koncepcją, która w mojej opinii, jest w stanie w istotny sposób udoskonalić wymienione koncepcje, jest hipoteza dopaminergicznego błędu predykcji nagrody (HDBPN). Zaproponowana w latach dziewięćdziesiątych XX wieku przez badaczy z *Salk Institute for Biological Studies* (Montague, 2006, s. 144) hipoteza pozwala sformułować teorię konkurencyjną w stosunku do tej sformułowanej przez Wegnera. Przyjmuje się, zgodnie z HDBPN, że wzorce wyładowań neuronów dopaminergicznych, obserwowane podczas warunkowania małp, zbieżne są z wartością błędu predykcji nagrody, będącego ważnym parametrem występującym w algorytmie TDRL (por. rozdział 3.), czyli jednej z metod uczenia maszynowego (Wolfram Schultz, Dayan, & Montague, 1997). Znaczy to, że wybrane zachowania zwierząt można opisać za pomocą wspomnianego algorytmu. Mechanizm uczenia się na podstawie wzmocnień wyjaśnia również inne zjawiska związane z układem dopaminergicznym, m.in. trudność w inicjowaniu ruchów w chorobie Parkinsona oraz wzorce zachowań zwierząt wystawionych na działanie substancji uzależniających. Sukcesywnie wzrasta zakres zjawisk, które wyjaśnia się przez przywołanie HDBPN. Współtwórca omawianej hipotezy, Read Montague, w pracy zatytułowanej *Why Choose This Book? How We Make Decisions* starał się wykazać, że HDBPN nie tylko odnosi się do zachowań uwarunkowanych biologicznie, ale – przy pewnym rozszerzeniu pojęcia nagrody – pozwala również wyjaśnić złożone zachowania ludzkie, w których istotny jest przede wszystkim czynnik przekonaniowy,

niezwiązany bezpośrednio z przetrwaniem i zachowaniem gatunku. Interpretacja Montague jest odważna i może zdawać się wysoce spekulatywna. Ta pierwsza cecha – śmiałość – wskazuje na heurystyczną płodność HDBPN, natomiast druga – spekulatywność – zostanie znacznie osłabiona, jeśli wykaże się, że mechanizm dopaminergicznego błędu predykcji nagrody (DBPN) jest składnikiem struktury działania intencjonalnego.

Według mojej wiedzy, dotychczas nie podjęto próby zintegrowania wyników zgromadzonych przez psychologów intencji z hipotezą dopaminergicznego błędu predykcji nagrody. Powód tego jest następujący: każdy z wymienionych nurtów badawczych skupia się na eksplorowaniu zjawisk z właściwego dla niego poziomu. Neurobiologowie odwołują się do wzorców wyładowań neuronów dopaminergicznych oraz konstruują wyjaśniające te zjawiska modele obliczeniowe, psychologowie z kolei poszukują głównie związków między czasowymi charakterystykami określonych potencjałów a reakcjami behawioralnymi, niekiedy uzupełniając wyniki badań empirycznych o raporty introspekcyjne, za pomocą których identyfikowali treści stanów intencjonalnych. Problemem do tej pory nierozwiązanym pozostaje kwestia teoretycznego powiązania zależności ustalonych dla każdego z poziomów z osobna. Pokazuje to jak ważnym zagadnieniem dla wskazanego wyżej przedsięwzięcia integrującego pozornie niezgodne podejścia jest wybór określonej ramy teoretycznej.

Współczesna kognitywistyka, jak każda dynamicznie rozwijająca się dziedzina nauki, dysponuje stosunkowo obszernym zestawem tego typu ram. Do najbardziej popularnych zalicza się: „koncepcję rozszerzonego umysłu, enaktywizm, predykcyjną teorię umysłu, teleosemantykę" (za: Miłkowski, 2015). Paradoks polega na tym, że wymienione podejścia w zasadzie nie są obecne w badaniach, które stawiają sobie za cel stworzenie zintegrowanej koncepcji działań intencjonalnych. Być może jest tak dlatego, że żadna z tych koncepcji nie oferuje własnego, oryginalnego ujęcia działania intencjonalnego. Nie znaczy to jednak, że poszukując ram teoretycznych, które mogłyby posłużyć jako spoiwo wiążące psychologię intencji z podejściami obliczeniowymi w rodzaju HDBPN, należy w ogóle zrezygnować z poszukiwania takich koncepcji. W mojej opinii, taką użyteczną ramę pojęciową, która dobrze nadaje się do zintegrowania danych neurobiologicznych z wynikami psychologii intencji, jest teoria intencjonalności Johna Searle'a, opublikowana w 1983 roku. Co prawda, nie funkcjonuje ona w głównym nurcie kognitywistyki, jednak wybrane jej elementy są przywoływane przez psychologów intencji. Zwykle korzystają oni

z wprowadzonego przez amerykańskiego filozofia podziału na prior intencję oraz intencję w działaniu (dystynkcja ta pozwala m.in. odróżnić przebieg zaplanowanych działań intencjonalnych od działań spontanicznych). Z kolei badacze stosujący podejście neurobliczeniowe na skutek Searle'owskiej krytyki obliczeniowej teorii umysłu (*computational theory of mind*) albo unikają daleko idących generalizacji w odniesieniu do proponowanych przez siebie modeli, albo poszukują rozwiązań, które pomogłyby przezwyciężyć wskazaną przez filozofa z Berkeley trudność (patrz: *value-based computational theory of mind* Reada Montague'a (Montague, 2006, s. 123)), która polega na niemożności wyjaśnienia ludzkich zdolności poznawczych w kategoriach operacji czysto syntaktycznych zdefiniowanych przez Alana Turinga (Searle, 2008a). Te nawiązania do propozycji Searle'a są jednak zbyt zdawkowe, aby mogły się stać podstawą koncepcji zintegrowanej struktury działania intencjonalnego. W niniejszej pracy przedstawię obszerniejszą analizę koncepcji intencjonalności Searle'a oraz pokażę, jak można wykorzystać niektóre z jej idei w konstruowaniu modelu działania intencjonalnego, uwzględniającego zarówno wyniki psychologii intencji jak i ustalenia podejścia neurooobliczeniowego.

Całościowe ujęcie działań intencjonalnych wymaga, aby zajmujący się nimi badacz odniósł się do dwóch istotnych problemów. Pierwszy, dotyczy charakterystyki działań złożonych i ich relacji do działań prostszych (patrz: problem rozciągnięcia konstruktywistycznej hipotezy Daniela Wegnera na przypadki działań złożonych). Drugi, związany jest ze znalezieniem sposobu integrowania danych, zebranych z różnych poziomów złożoności i wyrażonych w odmiennych aparatach pojęciowych. W przypadku niniejszej dysertacji jest to problem powiązania twierdzeń obliczeniowego modelu selekcji zachowań z wynikami badań psychologii intencji oraz z teorią intencjonalności Johna Searle'a. Problemy te uwzględnione są w głównej hipotezie pracy, którą sformułuję w dalszej części rozdziału.

Najpierw jednak przedstawię racje przemawiające za potrzebą stworzenia zintegrowanego modelu działań intencjonalnych.

1.3 O potrzebie multidyscyplinarnego podejścia do badania działań intencjonalnych

Podstawową racją uzasadniającą potrzebę stworzenia zintegrowanego modelu struktury działań intencjonalnych jest to, że w wielu różnych dyscyplinach naukowych (od filozofii przez psychologię, kognitywistykę, a na sztucznej inteligencji kończąc) uzyskano znaczące wyniki, które bezpośrednio lub pośrednio dotyczą takich działań, jednak brane z osobna wyjaśniają co najwyżej pewne ich fazy, aspekty lub cechy. W żadnej z tych dyscyplin nie tylko nie stworzono całościowego modelu działań intencjonalnych, ale nawet nie sformułowano satysfakcjonujących odpowiedzi na wiele fundamentalnych pytań dotyczących ich natury. Do najważniejszych z nich zaliczyć można:

1. Czy działania intencjonalne są zdeterminowane, czy też zawierają one składnik niezdeterminowany (problem wolnej woli, łac. *liberum arbitrium*) (Honderich, 2001)?
2. Czy sprawca działania zna jego rezultat (skutek) i dalsze konsekwencje? Kiedy i w jakim stopniu ponosi on odpowiedzialność za skutek, a kiedy za konsekwencje swojego działania (problem odpowiedzialności moralnej i karnej za podejmowane czyny) (Roskies, 2006)?
3. Czy zachowania, w których poświęca się własne życie przedkładając nad nie inne wartości (abstrakcyjne lub konkretne), wskazane przez określone normy (obyczajowe, religijne, prawne, itp.), dają się wyjaśnić za pomocą subtelnych mechanizmów biologicznych (problem wartości adaptacyjnej działań „zaniedbujących" potrzebę przetrwania) (Montague, 2006, s. 110)?
4. Na czym polega świadoma, skuteczna kontrola działań? Dlaczego niektórzy mają tzw. słabą wolę i są podatni na uzależnienia, a inni potrafią oprzeć się różnego rodzaju pragnieniom czy pokusom (Montague, 2006, s. 105)?
5. Czy reguły racjonalności to filozoficzno-ekonomiczne konstrukty o charakterze czysto instrumentalnym, czy raczej schematy zachowań reprezentujące realne „siły" kształtujące nasze faktyczne wybory i działania (Searle, 2001)?

Wymienione problemy są doniosłe nie tylko z teoretycznego punktu widzenia. Ich rozstrzygnięcia mogą znacząco wpłynąć na pojmowanie natury ludzkiej (wizja naturalistyczna człowieka *vs* wizja antynaturalistyczna), funkcji kary w systemie prawnym

(odpłata *vs* resocjalizacja) czy istoty moralności (utylitaryzm, deontologizm czy etyka cnót).

Każde z powyższych zagadnień ma, jak się wydaje, podobną strukturę. Z jednej strony odwołują się one do specyficznych dla swojej dziedziny argumentów i danych (w przypadku problemu wolnej woli jest to na przykład ontologia), z drugiej strony – implikują, mniej lub bardziej jawnie, pewne modele działań intencjonalnych (w kontekście problemu wolnej woli jest to np. model stworzony na podstawie introspekcyjnego doświadczenia, które ma uzasadnić przekonanie o nieuwarunkowanym, swobodnym wyborze jednej z dostępnych opcji). O ile wnioski dotyczące przedmiotu badań danego obszaru oparte są na ogół na rzetelnych analizach, argumentach i koncepcjach teoretycznych zgodnych z zasadami metodologicznymi danej dziedziny, o tyle część dotycząca cech oraz funkcjonowania działań intencjonalnych rzadko uwzględnia najnowsze wyniki badań naukowych. Dostępne dane zazwyczaj są dobierane w sposób wybiórczy (na przykład, w kontekście problemu wolnej woli najczęściej przywoływany jest eksperyment Libeta), a opisy mechanizmów organizujących działania nie wychodzą poza podstawowe intuicje.

Zaproponowana w niniejszej pracy analiza wpływu stanów intencjonalnych na dobór zachowań służy przezwyciężeniu wspomnianego ograniczenia poprzez połączenie rezultatów badań z dwóch niezależnie rozwijających się nurtów badawczych: psychologii intencji oraz tzw. neurobiologicznych podstaw procesów decyzyjnych. Uwzględnienie najważniejszych wyników obu koncepcji pozwala, moim zdaniem, wyjść poza „lokalne optima", czyli hipotezy poprawnie wyjaśniające tylko wybrane zjawiska z danego obszaru, wyodrębnione ze względu na potrzeby konkretnej perspektywy badawczej. Kiedy jednak uwzględnieni się także wiedzę z innej dyscypliny naukowej i włączy do rozważań nowe, zgromadzone w niej dane, wtedy okazuje się, że wyjaśnienie uznawane wcześniej za poprawne jest adekwatne tylko w ograniczonym, wąskim zakresie. Zintegrowany model działań intencjonalnych, będący rozwinięciem głównej hipotezy dysertacji, umożliwia wgląd w kluczowe mechanizmy decydujące o zdolności do realizowania złożonych sekwencji zachowań. W efekcie, uwzględnienie ZMDI pokazuje, że dotychczasowe, ugruntowane często w doświadczeniu introspekcyjnym, pojmowanie działań intencjonalnych nie dość, że jest nader uproszczone, to jeszcze nie uwzględnia ważnego dla ich charakterystyki, aktualnego stanu wiedzy.

1.4 Teza

W mojej pracy stawiam i uzasadniam następującą tezę:

> **Złożone działanie intencjonalne to system dwóch współdziałających mechanizmów: (i) uczenia się ze wzmacnianiem oraz (ii) planowania na podstawie wiedzy zgromadzonej w formie sieci stanów intencjonalnych.**

Mechanizm uczenia się ze wzmacnianiem odpowiada za: (1) zaspokajanie podstawowych potrzeb agenta, które pojawiają się już we wczesnej fazie rozwoju, (2) uczenie się nowych, bardziej złożonych zachowań oraz (3) generowanie tzw. elementarnych warunków spełniania, czyli podstawowych jednostek umożliwiających tworzenie stanów intencjonalnych. **Planowanie** odpowiada za optymalizację mechanizmu uczenia się ze wzmacnianiem – poprzez włączenie w jego funkcjonowanie nowych typów nagród oraz wiedzy dziedzinowej. Obydwa mechanizmy, choć korzystają ze specyficznych dla siebie typów reprezentacji, wpływają na siebie, zachowując przy tym autonomię w zakresie sposobów działania. Wymienione mechanizmy – w swej w pełni rozwiniętej postaci – tworzą zhierarchizowaną strukturę, w której proces planowania „kształtuje" przebieg procesu uczenia się ze wzmacnianiem.

„Nagroda" oraz „planowanie" są pojęciami, które zostały tu użyte w następującym znaczeniu:

- **Nagroda** to informacja pozyskiwana przez agenta w trakcie jego interakcji ze środowiskiem i mająca dla niego określoną wartość. W języku potocznym pojęcie „nagrody" odnosi się do sytuacji lub obiektów waloryzowanych pozytywnie. Tutaj przyjmuję jednak konwencję powszechnie respektowaną w neuronauce obliczeniowej (Antonio Rangel i in., 2008; Cabanac, 1992), że informacja będąca nagrodą dla agenta może mieć zarówno wartość pozytywną, jak i negatywną. To, co potocznie nazywa się karą, będzie tu określane jako nagroda ujemna. W przypadku organizmów biologicznych detekcja wartościowej informacji odbywa się za pomocą układu nerwowego wyposażonego w określone dyspozycje. Na poziomie funkcjonalnym poszczególne typy nagród (mogą nimi być reprezentacje – neuronalne lub umysłowe – obiektów z otoczenia, zachowań innych osobników, a także własnych stanów danego organizmu) wywołują wrodzone lub nabyte reakcje

organizmu na dany rodzaj reprezentacji. Tak rozumiana nagroda jest zasadniczym składnikiem mechanizmu uczenia się.

- **Planowanie** utożsamiane jest z szeroko pojętym procesem deliberacyjnym, rozumianym jako zespół procesów poznawczych poprzedzających podjęcie decyzji. W procesie planowania wykorzystywane są różne typy rozumowań, tworzone alternatywne scenariusze realizacji celu, wartościowane dostępne opcje, itp.

W konstruowaniu tezy dysertacji oraz powiązanego z nią Zintegrowanego Modelu Działań Intencjonalnych (ZMDI) inspiracje, a w niektórych przypadkach również gotowe rozwiązania, czerpałem z teoretycznych koncepcji sformułowanych w trzech różnych obszarach badawczych. Są to:

1. teoria intencjonalności Johna Searle'a,
2. koncepcje teoretyczne i powiązane z nimi empiryczne badania psychologiczne dotyczące prostych działań intencjonalnych,
3. model obliczeniowy uczenia się ze wzmacnianiem wykorzystany w badaniach nad neurobiologicznymi podstawami procesów decyzyjnych oraz nad kontrolą zachowań.

Przedstawione poniżej objaśnienia, które się odnoszą do ZMDI i jego składników, można potraktować jako rozbudowaną argumentację na rzecz postulowanej przeze mnie tezy.

1.5 Cel i metoda

Głównym problemem, który należy rozwiązać, zanim przystąpi się do konstrukcji ZMDI, jest ustalenie natury relacji istniejących między mechanizmem uczenia się ze wzmacnianiem a planowaniem. Dlatego też ważnym celem pracy jest doprecyzowanie struktury obu wymienionych form kontroli zachowań.

Po wykonaniu tego zadania będzie można przystąpić do konstrukcji ZMDI. W trakcie tej pracy poddam ocenie wskazane w punkcie 1.4 źródła inspiracji po to, aby zdecydować, które ich składniki można włączyć w strukturę ZMDI. Kryteriami, którymi będę się kierował w tej ocenie, są: z jednej strony efektywność eksplanacyjna rozważanych koncepcji, a z drugiej – stopień, w jakim dane rozwiązanie daje się uzgodnić z innymi składnikami modelu (kryterium koherencyjne).

1.6 Zintegrowany model działań intencjonalnych (ZMDI)

Jak sygnalizowałem wyżej, zintegrowany model działań intencjonalnych pokazuje, jak w ramach jednego systemu, decydującego o podjęciu działania, współdziałają dwa podsystemy: rdzeniem tego pierwszego jest mechanizm uczenia się ze wzmacnianiem, rdzeniem tego drugiego jest mechanizm planowania. Aby obydwa mogły działać w sposób niezakłócony, agent powinien być wyposażony w następujące funkcje oraz implementujące je układy:

- podsystem odbioru, przetwarzania i ewaluacji informacji pochodzących ze środowiska,
- podsystem monitorowania i aktywacji celów dostosowanych do aktualnego stanu świata oraz wewnętrznego stanu agenta uwzględniający zmiany stanu otoczenia (P. Cichosz, 2007, s. 715) oraz zmiany stanów (zarówno cielesnych, jak i umysłowych) samego agenta (patrz: m.in. problem utrzymania homeostazy),
- podsystem generujący sieć stanów intencjonalnych, czyli formy reprezentowania aktualnych oraz przyszłych stanów rzeczywistości, w tym zachowań innych agentów.

Proponowany przeze mnie model (patrz: Diagram 1 poniżej) wyjaśnia złożone przypadki działań intencjonalnych, czyli działań składających się z komponentu umysłowego (w jego skład wchodzi w pierwszej kolejności intencja, czyli zamiar) oraz zachowaniowego (np. motorycznego)[9]. Model ten posiada potencjał eksplanacyjny, gdyż sam mechanizm uczenia się ze wzmacnianiem pozwala zrealizować dowolnie wybrany cel behawioralny w sposób niemal optymalny - o ile założy się, że proces uczenia się agenta jest nieskończony, że dysponuje on nieograniczonymi zasobami energetycznymi oraz odpowiednio bogatym zbiorem reprezentacji nagród. Tego typu wyidealizowany model umożliwia wyjaśnienie nie tylko stosunkowo prostych przypadków uczenia się odruchów warunkowych, ale także, wbrew obiegowej opinii, złożonych umiejętności oraz celów wymagających koordynacji wielu zachowań. Efekt ten jest możliwy na skutek połączenia dwóch elementów, które wspólnie tworzą algorytm efektywnego selekcjonowania

[9] Za podstawę teoretyczno-empiryczną przyjętej konstrukcji przyjąłem wyniki aktualnych badań neurobiologicznych nad mózgowymi korelatami procesów decyzyjnych (hipoteza dopanergicznego błędu predykcji nagrody oraz prace informatyków dotyczące obliczeniowych podstaw uczenia się ze wzmacnianiem (Sutton, 1998)). Z kolei w konstrukcji mechanizmu planowania wykorzystałem ustalenia teorii intencjonalności Johna Searle'a oraz wyniki najnowszych badań z obszaru uczenia maszynowego, które dotyczą prób wzbogacenia metody uczenia się ze wzmacnianiem o wiedzę domenową oraz elementy planowania.

zachowań. Pierwszy z nich odnosi się do zdolności reprezentowania oraz ewaluacji informacji czerpanych z szeroko pojętego otoczenia, obejmującego zarówno środowisko zewnętrze organizmu, jak i jego własne stany cielesne i umysłowe. Agent – na podstawie wskazanych reprezentacji – nie tylko wie, w jakim stanie świata się obecnie znajduje, ale również zna konsekwencje tego faktu, tzn. zna błąd predykcji nagrody. Drugi element odnosi się do „wbudowanych" w proces przetwarzania informacji funkcji uczenia się i optymalizacji zachowań, która umożliwia agentowi przejście od losowej eksploracji, opartej na metodzie prób i błędów, do eksploatacji, czyli wyboru działań dostosowanych do bieżącego kontekstu środowiskowego. Tego rodzaju wybór korzysta z pozyskanych wcześniej doświadczeń.

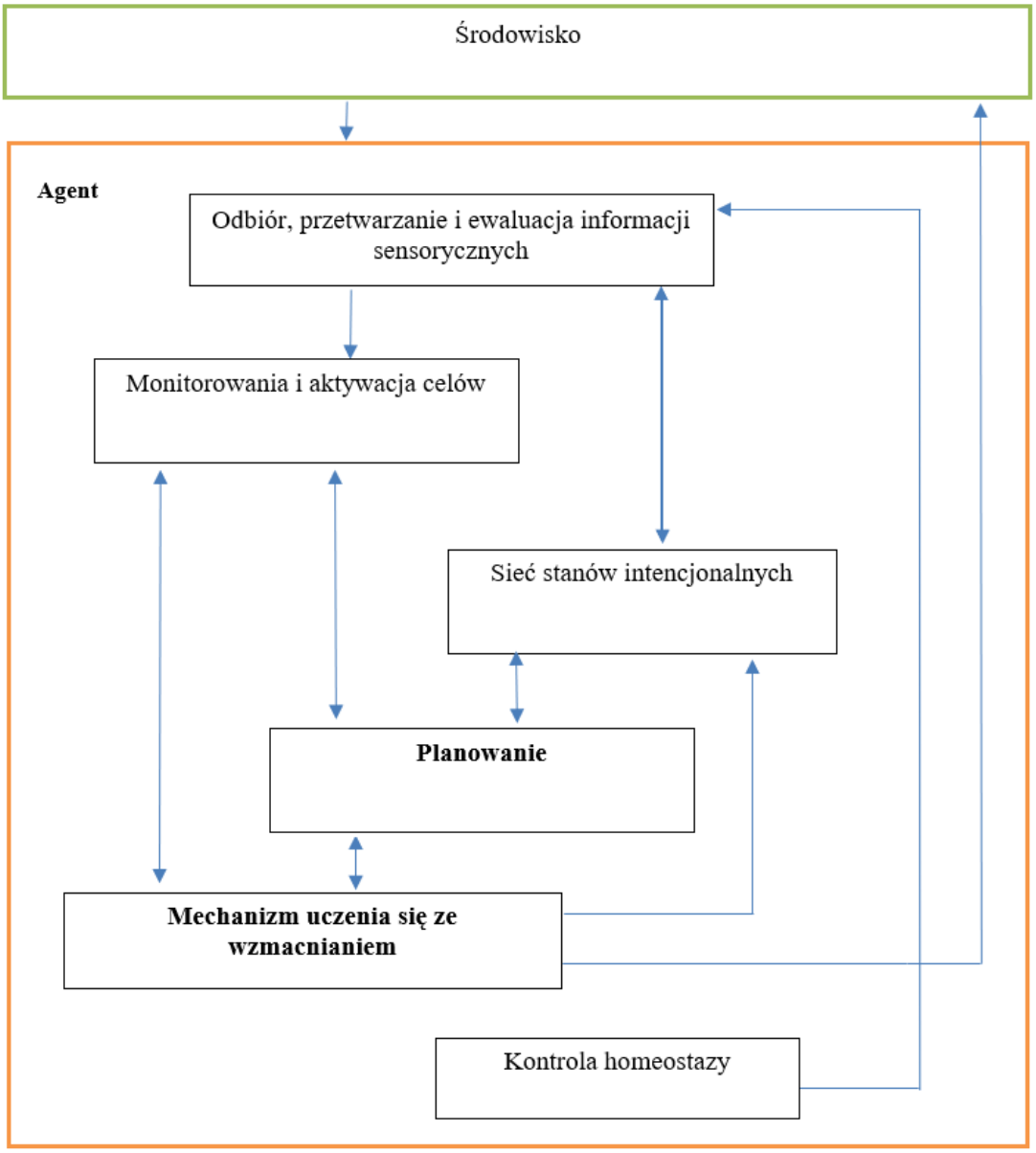


Diagram 1. Zintegrowany model działań intencjonalnych (ZMDI) – ujęcie wysokopoziomowe.

Oczywiste staje się, że model z takimi założeniami idealizującymi jest wysoce nierealistyczny i odniesienie go do faktycznych sytuacji wymaga skonkretyzowania założeń, Uważam jednak, że dopiero za sprawą tak dalece posuniętej idealizacji odsłonić można najważniejsze składowe działania intencjonalnego. Po zniesieniu lub osłabieniu założeń można już stosować urealistyczniony model do wyjaśniania faktycznych zachowań będących efektem działania mechanizmu uczenia się ze wzmacnianiem. Chodzi tu przede wszystkim o uwzględnienie tego, że zwierzęta, w tym ludzie, nie dysponują nieograniczonymi zasobami energetycznymi i nieograniczonym czasem, a środowisko, w którym funkcjonują, jest niezwykle „wymagające" (jak wiadomo, pełni funkcję selektora w mechanizmie doboru naturalnego). W tej sytuacji nasuwają się dwa rozwiązania. Pierwsze z nich polega na ograniczeniu zakresu dających się wyuczyć zachowań do – nabywanych w trybie dziedziczenia – „gotowych" do zastosowania wzorców (selekcja oraz transmisja repertuaru zachowań odbywa się w tym przypadku za pośrednictwem genów). Mechanizm uczenia się ze wzmacnianiem nadal okazuje się niezbędny do przetrwania, jednak nie musi się on już „troszczyć" o wykształcenie podstawowych zachowań adekwatnych do wymagań środowiska. Proces uczenia sprowadza się do eksploracji niszy ekologicznej zajmowanej przez dany gatunek i do koordynacji wrodzonych zachowań. W tym przypadku, dziedziczone są również mechanizmy rozpoznawania nagród. Znaczy to, że waloryzowane pozytywnie lub negatywnie rodzaje pokarmów i płynów, korzystne albo niekorzystne stany środowiska bądź przyjazne albo wrogie zachowania innych agentów są w dużym stopniu określone z góry. Innymi słowy, różnorodność relewantnych dla agenta stanów jest zdeterminowana przede wszystkim przez jego potrzeby biologiczne, choć, jak widać na przykładach zachowań zwierząt wychowywanych w niewoli, istnieje w tym obszarze również otwartość na nowe możliwości (patrz: przypadek bonobo Kanzi (Segerdahl i in., 2005)). W takich sytuacjach można oczekiwać, że zwierzę będzie od pierwszych dni po urodzeniu sprawnie realizować specyficzne dla danego gatunku zachowania, a proces nauki dotyczył będzie głównie ich efektywnego zastosowania w danym kontekście środowiskowym. Jeśli pomiędzy środowiskiem a potrzebami i repertuarem zachowań agenta następowało będzie „dopasowanie", wówczas w bardzo krótkim czasie zwierzę uzyska pełną zdolność funkcjonowania w dostępnej dla niego niszy. Jeśli jednak okaże się, że środowisko stawia przed młodym organizmem zbyt wysokie wymagania, to jego byt będzie zagrożony.

Drugie rozwiązanie, urealniające wyidealizowany model polega na modyfikacji metody uczenia się ze wzmacnianiem. Sprowadza się to do wprowadzenia dodatkowej informacji w mechanizmy sekwencjonowania zachowań oraz uczenia się. Tego typu nowa informacja pozwala agentowi wykroczyć poza dostępne obserwacje oraz nagrody (w żargonie specjalistów uczenia maszynowego mówi się często o przezwyciężeniu ograniczeń wynikających z tzw. własności Markowa[10]). W tym przypadku sprawne funkcjonowanie zaraz po urodzeniu nie wymaga już dziedziczenia złożonych zachowań. Nabywane są one stopniowo w procesie uczenia się. Nadmiar odziedziczonych, gotowych do wykorzystania dyspozycji byłby balastem dla agenta, krępującym jego swobodę w kształtowaniu umiejętności dostosowanych do zmieniających się wyzwań środowiska oraz do pojawiających się coraz nowszych potrzeb.

W dysertacji proponuję następujące konkretyzacje mechanizmu uczenia się ze wzmacnianiem:

- Dołączenie do mechanizmu uczenia się ze wzmacnianiem zdolności do „konstruowania" nowych typów nagród (patrz: tzw. hipoteza nad-mocy Montague'a). W ten sposób organizm zyskuje swobodę w „decydowaniu" o tym, co jest dla niego wartościowe, a co nie. W szczególnym przypadku, abstrakcyjne stany intencjonalne, takie jak np. pragnienie sławy, oczekiwanie uznania za osiągnięcia (artystyczne, naukowe, itp.) czy chęć obrony wyznawanych wartości mogą uzyskać status nagrody. Wszystkie one mogą (lecz nie muszą) być w zasadniczy sposób oderwane od zdeterminowanych genetycznie nagród biologicznych.
- Poszerzenie repertuaru nagród o tzw. „nagrodę kształtującą", czyli o wiedzę dziedzinową wyrażoną w formie wartości skalarnych. Uwzględnienie takiej nagrody znacząco skraca czas potrzebny na „odkrycie" niemal optymalnej, dla osiągnięcia wybranego celu, sekwencji zachowań.
- Poszerzenie repertuaru sprawności agenta o umiejętność hierarchizowania zachowań. W ten sposób, podsystem planowania, współpracujący z mechanizmem uczenia się ze wzmacnianiem, pozwala na projektowanie złożonych celów (tzw. planów prowizorycznych), wymagających koordynacji wielu wysokopoziomowych zachowań oraz dostarczania nagród, odnoszących się do określonych stanów

[10] Własność Markowa nakłada na model środowiska wymóg, by decyzje dotyczące wyboru działania w danym stanie świata zależały jedynie od dostępnych w nim informacji. Czasami własność tę definiuje się jako niezależność procesu decyzyjnego od historii (tzw. proces bez pamięci).

intencjonalnych. Tego typu plany są kluczowym narzędziem optymalizacji procesu uczenia się, pozwalają one nie tylko rozciągnąć horyzont działań agenta, ale również minimalizować liczbę błędów popełnianych w trakcie eksploracji środowiska.

Każde z wymienionych rozszerzeń zakłada tworzenie nowych typów reprezentacji oraz istnienie nowych procesów ich przetwarzania. Tak urealistyczniony obraz działania mechanizmu uczenia się ze wzmacnianiem u członków podgatunku *homo sapiens sapiens* lepiej wyjaśnia generowanie złożonych zachowań intencjonalnych, niż omówione wyżej ujęcie odwołujące się wyłącznie do ugruntowanych ewolucyjnie procesów dziedziczenia. Poszerzony we wskazany wyżej sposób system zwiększa swoje możliwości adaptacyjne w porównaniu z systemem wykorzystującym wyłącznie transmisję repertuaru zachowań za pomocą genów. W obu przypadkach agent skutecznie dostosowuje się do wymagań niszy ekologicznej, jednak z perspektywy gatunku oraz jego możliwości przystosowawczych zmiana jest zasadnicza. Rozszerzony mechanizm uczenia się ze wzmacnianie jest bardziej ogólny i elastyczny, pozwala skutecznie działać w odmiennych typach środowiska (gorący klimat afrykański vs. ekstremalnie zimny klimat Syberii) oraz adaptować się do nagłych i drastycznych zmian, jak choćby klęski żywiołowe. Nadal jednak wszystkie umiejętności nabywane przez agenta są pochodną mechanizmu uczenia się ze wzmacnianiem, który nie obejmuje zdolności do planowania działań, dlatego kompletny model działań intencjonalnych wymaga kolejnych uzupełnień, w szczególności o bardziej złożone formy uczenia się oraz reprezentowania świata.

Przedstawione rozszerzenia metody uczenia się ze wzmacnianiem pokazują, jak istotną rolę w procesach adaptacyjnych pełnią zaawansowane systemy projektowania, wyboru i ewaluacji celów. Wyraźnie widać to na przykładzie osiągnięć reprezentantów gatunku *homo sapiens*, którzy na tle zwierząt z innych gatunków wyróżniają się niezwykłą wręcz elastycznością w dostosowywaniu się do skrajnie odmiennych typów środowiska. „Ceną", jaką przychodzi nam płacić za tę zdolność, jest znacząco dłuższy cykl rozwoju osobniczego, niż w przypadku zwierząt z innych gatunków, w tym naczelnych. Kiedy porówna się swoistą „nieporadność" niemowlęcia i złożone działania osób dorosłych, w szczególności specjalnie wyszkolonych ekspertów (np. zawodowych koszykarzy lub lekarzy), można by dojść do wniosku, że niemowlę i dorosły wyposażeni są w dwa, całkowicie różne systemy kontroli zachowań. Konkluzja taka nie jest jednak trafna, gdyż potrafimy dostatecznie precyzyjnie prześledzić, jak stan dojrzały stopniowo wyłania się ze stanu początkowego. Pozorna odmienność dwóch systemów organizacji działań staje się

widoczna, kiedy zauważymy, że mechanizm uczenia się ze wzmacnianiem dostępny jest w pełni już w chwili narodzin, natomiast mechanizm planowania, ze względu na swą silną zależność od sieci stanów intencjonalnych, jest wtedy jeszcze nieaktywny. Dopiero po latach praktyki mechanizm uczenia się ze wzmacnianiem jest stopniowo rozszerzany o planowanie. Procesy tworzenia nowych reprezentacji, które pozwalają zastosować planowanie oraz uczenie się ze wzmacnianiem do coraz bardziej złożonych celów, są niewątpliwie ważnymi czynnikami wpływającymi na rozwój całego systemu kontroli zachowań. W zaproponowanym modelu przyjęto, że za pojawienie się nowych typów reprezentacji odpowiedzialne są dwa główne mechanizmy.

Pierwszy to zdolność do generalizacji, wbudowana w podsystem „odbioru, przetwarzania i ewaluacji informacji sensorycznej" (Friston, 2003) oraz podsystem „zarządzania siecią stanów intencjonalnych" (Sutton i in., 1999). Każdy z nich służy do szukania wzorców (sensorycznych lub zachowaniowych), które pozwolą lepiej orientować się w środowisku lub bardziej efektywnie na nie oddziaływać. Przyjąłem, za Karlem Fristonem, że w przypadku podsystemu sensorycznego proces generalizacji polega w głównej mierze na kodowaniu predykcyjnym, zapewniającym zdolność rozpoznawania obiektów niezależnie od kontekstu, w którym się pojawiają. Bez względu na to, czy obiekt się porusza, czy jest w spoczynku, czy jest w pełni widoczny, czy częściowo przesłonięty, modele generatywno-rozpoznawcze, implementujące mechanizm kodowania predykcyjnego, zapewnią, że obiekt zostanie prawidłowo rozpoznany. Podobny cel „stawia sobie" podsystem zarządzania siecią stanów intencjonalnych w odniesieniu do zachowań. Z zapamiętanych ich sekwencji (tzw. taktyki (*policy*)) próbuje on wyłowić – niezależne od celu oraz typu nagrody – wzorce cechujące się odpowiednią uniwersalnością. Znalezione wzorce umożliwiają, ważne z perspektywy planowania, operacje takie jak: tworzenie hierarchii zachowań czy „transfer" umiejętności między różnymi sytuacjami pojawiającymi się w odniesieniu do realizowanych celów. Specjaliści od uczenia maszynowego nazywają tego typu wzorce „opcjami" (Sutton i in., 1999; Yao i in., 2014), z kolei John Searle używa kategorii „umiejętności tła" lub tzw. „działania podstawowego" (Searle, 1983, s. 100). Tak wyznaczone zachowania wyższego rzędu stają się częścią repertuaru wzorców zachowań, wykorzystywanych przez rozszerzony mechanizm uczenia się ze wzmacnianiem, w istotny sposób zwiększając jego efektywność. Czasochłonna i kosztowna energetycznie eksploracja zostaje poważnie ograniczona.

Trudno wyobrazić sobie model działania intencjonalnego, w którym pominięty zostałby moduł zawierający mechanizm konstruowania sieci stanów intencjonalnych. Znane mi wyniki badań naukowych (Jodzio, 2008; Nęcka i in., 2006a) nie dostarczyły do tej pory wiedzy, która pozwalałaby na szczegółowy opis mechanizmów odpowiedzialnych za konstrukcję tego typu sieci oraz tworzących ją składników, takich jak: pragnienia, przekonania, intencje, obawy, wahania, itp. Skoro nie można sięgnąć do ustaleń nauk empirycznych, to należy szukać wsparcia w tych dziedzinach, w których prowadzi się racjonalną i rygorystyczną refleksję nad działaniami intencjonalnymi. Zakładam, że szkic takiego mechanizmu zawiera zaproponowana przez Johna Searle'a teoria intencjonalności, w której stan intencjonalny – S(p) – to odpowiednio zorganizowany zbiór warunków spełniania. W pracy wykorzystuję wybrane elementy teorii intencjonalności Searle'a, abstrahując od ogólnej struktury pojęciowej, w którą zostały wpisane. Przyjmuję mianowicie, że warunki spełniania wyznaczające treść intencji, przekonań, pragnień oraz innych stanów intencjonalnych opierają się na reprezentacjach wykorzystywanych przez metodę uczenia się ze wzmacnianiem. Innymi słowy, postuluję wprowadzenie do modelu mechanizmu generowania hierarchii reprezentacji, u którego podstaw leży operacja pozwalająca agentowi tworzyć reprezentacje złożone z reprezentacji prostszych, wykorzystywanych m.in. przez algorytm RL (*reinforcement learning*). Algorytm implementujący mechanizm uczenia się ze wzmacnianiem pozwala formalnie zdefiniować tego typu zależność w następujący sposób:

$$s_t, z_t / Z_t \rightarrow s_{t+1}, \delta_{t+1}, r_{t+1}, o_{t+1}, O_{t+1}$$

Poszczególne symbole odnoszą się do następujących reprezentacji:

- s_t, s_{t+1} – reprezentacja stanu świata przed (s_t) oraz po (s_{t+1}) realizacji zachowania z_t,
- z_t – reprezentacja elementarnego zachowania (wrodzonych odruchów, prostych, wyuczonych ruchów cielesnych, a także artykulacji, czyli wyuczonych sposobów wytwarzania dźwięków mowy) zrealizowanego w chwili t; odpowiednio zorganizowana sekwencja z_t tworzy zachowanie złożone Z,
- δ_{t+1} – reprezentacja błędu predykcji nagrody po wykonaniu zachowania z lub Z,
- r_{t+1} – reprezentacja nagrody pozyskanej po wykonaniu zachowania z lub Z,
- o/O_{t+1} – elementarna lub uogólniona obserwacja środowiska w chwili t+1, na podstawie której wyznaczany jest stan środowiska, w którym znalazł się agent.

Zapisanej powyżej formuły nie należy interpretować czysto obiektywistycznie jako opisu związków między faktycznymi stanami świata, zachowaniami czy nagrodami. Formuła ta odnosi się do odpowiadających im reprezentacji, poprzez które są one dostępne („jawią się") agentowi. Choć obiekty te dane są w sposób zapośredniczony, to agent traktuje je jako należące do realnego świata i projektuje swoje działania ze względu na ich fizyczne cechy dane poprzez treść reprezentacji.

Przedstawiona wyżej formuła opisuje związek pomiędzy zachowaniem agenta w danym stanie świata a skutkiem tego zachowania, jakim jest pojawienie się nowego, zmienionego w stosunku do poprzedniego, stanu świata. Podsystem konstruujący stany intencjonalne zaczyna w ten sposób dysponować bogatym zbiorem reprezentacji, które może zacząć organizować w stany (reprezentacje złożone), a docelowo również w sieć stanów (Searle, 1983, s. 174). Jest to możliwe przy założeniu, że każdy, nawet najprostszy cel wymaga koordynacji bardzo wielu niskopoziomowych zachowań (np. ruchów).

Gdy połączy się efekty działania (1) mechanizmu uczenia się ze wzmacnianiem, (2) mechanizmu planowania, (3) procesu generalizacji oraz (4) generowania sieci stanów intencjonalnych, to uzyskujemy układ o architekturze samowspornej (*bootstrap*). Dysponujący nim organizm posiada predyspozycje niezbędne do tego, by przekształcić zbiór prostych zachowań w złożone celowe sekwencje, a następnie – by wykształcić na ich podstawie hierarchię zachowań oraz utworzyć sieć stanów intencjonalnych „zasilaną" warunkami spełniania (*conditions of satisfactions*), które powstają w trakcie interakcji między agentem a środowiskiem.

Kontrola zachowań nawet w swej najbardziej zaawansowanej postaci odbywa się z wykorzystaniem obu mechanizmów: uczenia się ze wzmacnianiem i planowania. Jak twierdzi Read Montague (Montague, 2006, s. 97), precyzyjne planowanie celów byłoby nie tylko zawodne, ale również bardzo kosztowne, trudno bowiem przewidzieć i przeanalizować wszystkie zdarzenia, które mogą pojawić się w środowisku naturalnym. W tej sytuacji optymalny wydaje się system, który opracuje i wyznaczy jedynie najważniejsze elementy niezbędne do osiągnięcia celu, a szczegóły wykonawcze „deleguje" do podsystemu uczenia się ze wzmacnianiem, który dobierze odpowiednie zachowania niezbędne do zrealizowania wybranego działania. Tak przebiega na wysokim poziomie ogólności „podział kompetencji" w proponowanym w dysertacji modelu działań intencjonalnych. Współpraca pomiędzy wymienionymi mechanizmami (uczeniem się i

planowaniem) zapewniona jest przez: (1) nagrody biologiczne, ale także i abstrakcyjne (mogą to być stany świata, idee, pragnienia) decydujące o tym, do czego agent będzie dążył albo czego będzie unikał (zgodnie z przyjętą konwencją nagrody mogą mieć wartość ujemną), gdyż postrzegać je będzie jako pożądane lub niepożądane z perspektywy przyjętych przez niego kryteriów wartościowania, (2) „nagrody kształtujące" odpowiedzialne za dostarczanie wiedzy domenowej do mechanizmu sekwencjonującego zachowania oraz (3) zachowania wyższego rzędu umożliwiające agentowi wielokrotne stosowanie nabytych umiejętności w nowych okolicznościach – bez konieczności powtarzania procesu uczenia. Za pomocą wymienionych czynników podsystem planowania jest w stanie „wysterować" podsystem uczenia się ze wzmacnianiem nie naruszając jego zasad działania, w szczególności zdolności do poszukiwania optymalnej strategii doboru zachowań. By jednak obydwa mechanizmy mogły efektywnie kooperować, potrzebny jest jeszcze jeden ważny składnik – prior intencja.

Szczególny stan intencjonalny, którym jest prior intencja, pełni w zaproponowanym modelu dwie istotne funkcje. Pierwsza z nich polega na tym, że ma ona dostęp do podsieci stanów intencjonalnych, reprezentującej plan utworzony w procesie deliberacji. W tym przypadku prior intencja to nic innego, jak „wskaźnik"[11] podsieci stanów intencjonalnych. Pozwala ona włączyć dany plan realizacji celu w sieć przekonań podmiotu, powiązać z innymi planami oraz pragnieniami. W związku z tym, treść prior intencji może być bardzo złożona i analizowana na wielu poziomach opisu (patrz: efekt akordeonu przeanalizowany przez Searle'a).[12] Opracowanie planu oraz wyznaczenie prior intencji jest warunkiem koniecznym, by dane działanie uznane zostało za zaplanowane. Posiadanie planu nie wystarcza jednak do podjęcia działania. Za ten aspekt funkcjonowania prior intencji odpowiada podsystem monitorowania i realizacji celów. Zajmuje się on monitorowaniem napływających informacji z perspektywy ich doniosłości dla realizacji nadzorowanych przez niego celów. Gdy w dostępnym strumieniu danych rozpoznana zostanie informacja relewantna dla jednego z nadzorowanych celów, wówczas podsystem ten „decyduje" – czy podjąć decyzję o aktywacji tego celu i równoczesnej dezaktywacji celu aktualnie realizowanego, czy też informację tę w tym momencie zignorować. Efektywne działanie

[11] „Wskaźnik" to polskie tłumaczenie angielskiego słowa „pointer". W językach programowania, np. w języku C / C++, wskaźnik to typ zmiennej przechowującej adres określonego obszaru pamięci, w którym mogą znajdować się m.in. dane programu. W zaproponowanym w pracy modelu przyjęto, że prior intencja pełni podobną funkcję w podsystemie monitorowania celów, stąd użycie słowa wskaźnik.

[12] Szczegółowy opis efektu akordeonu znajduje się w rozdziale 2, w punkcie dotyczącym złożonych działań intencjonalnych.

podsystemu monitorowania wymaga, aby prior intencja została do niego włączona w chwili, gdy zostanie ona utworzona.

Aby działanie intencjonalne mogło zostać podjęte, inicjujący je zamiar musi zostać skonkretyzowany do postaci osiągalnego – z perspektywy agenta – stanu rzeczy. Ten stan to cel, który podlega monitorowaniu przez wyspecjalizowany podsystem, który m. in. ustala – czy występujące aktualnie okoliczności pozwalają na aktywację celu. Jeśli napływające dane wskazują, że dostatecznie wysoka jest szansa na jego osiągnięcie, to odpowiadający prior intencji plan zostaje aktywowany i rozpoczyna się proces jego realizacji. W zaproponowanym ujęciu przyjmuje się, że plan obejmuje jedynie najważniejsze działania składowe oraz powiązane z nimi podcele. Niezbędne, niskopoziomowe zachowania wyselekcjonowane zostają przez podsystem uczenia się ze wzmacnianiem. Wskazany schemat działania modelu jest przejawem kontroli typu *top-down*, w którym wysokopoziomowe cele w złożonych stanach intencjonalnych „sterują" mechanizmami i strukturami niższego poziomu takimi, jak na przykład skurcze mięśni, choć równocześnie nie determinują ich wprost. Zakłada się, że każdy z podsystemów decydujących o kontroli zachowań funkcjonuje niezależnie, tzn. działa na podstawie odrębnych reguł, choć jednocześnie otwarty jest na informacje/reprezentacje pochodzące z podsystemu z nim współpracującego.

Przedstawione powyżej najważniejsze cechy modelu złożonych działań intencjonalnych nie uwzględniają wielu istotnych aspektów tej formy inteligentnego działania. Proponowane tu ujęcie, jak zaznaczyłem wcześniej, ma charakter modelowy, a więc dąży do uchwycenia najbardziej istotnych cech działań intencjonalnych. Znaczy to, że proponowana konstrukcja ma charakter wyidealizowany. Dopiero po jej konkretyzacji, tj. rozwinięciu wielu jej elementów i odniesieniu ich do obszerniejszej klasy prowadzonych w tym zakresie badań, otrzymamy bardziej realistyczny model złożonego działania intencjonalnego. Zadanie to zrealizowane zostało w ostatnim rozdziale pracy. Przedstawiony we wprowadzeniu model, choć pomija szereg szczegółów, w mojej opinii, dobrze przybliża istotę złożonego, czyli typowego dla istot ludzkich, działania intencjonalnego. Tworzące je składniki oraz funkcjonujące między nimi zależności wyraźnie pokazują, że nie da się stworzyć takiego modelu bez uwzględnienia wiedzy – zarówno teoretycznej, jak i empirycznej – wypracowanej w różnych dyscyplinach naukowych. Na tle złożonego modelu działań intencjonalnych widać m.in. dlaczego oryginalna i płodna poznawczo propozycja Reada Montague'a, traktująca uczenie się ze

wzmacnianiem jako główny mechanizm realizacji celów, nie wystarcza do wyjaśnienia najbardziej złożonych aktywności angażujących procesy deliberacji i planowania. Zaproponowany przeze mnie model zawiera również nowe rozumienie poczucia sprawstwa (*sense of agency*). Propozycja ta wyrosła z krytyki kontrowersyjnej koncepcji sprawstwa przedstawionej przez Daniela Wegnera. W proponowanym w pracy modelu poczucie sprawstwa jest przejawem procesów reprezentacjotwórczych, a nie – jak u Wegnera – rezultatem procesu konstruktywistycznego, mającego uspójnić działania z takimi potrzebami, jak bycie w pełni racjonalnym agentem. W opracowanej przeze mnie koncepcji to agent, z pomocą reprezentacji, stopniowo „buduje" złożoną sieć stanów intencjonalnych, która z czasem pozwala mu aktywnie kształtować otoczenie poprzez długoterminowe planowanie. Opracowana konstrukcja pozwala też osadzić wybrane elementy teorii intencjonalności Searle'a w kontekście dobrze ugruntowanych, współczesnych wyników badań empirycznych i teoretycznych, w tym modeli obliczeniowych. Oparcie kontroli zachowań na niskopoziomowym mechanizmie uczenia się ze wzmacnianiem oraz na planowaniu pozwala nie tylko wyjaśnić proces powstawania złożonych działań na bazie prostych odruchów, ale również odpowiedzieć, przynajmniej częściowo, w jaki sposób przebiega proces stopniowej automatyzacji działań rutynowych. Wymienione rezultaty oraz szczegółowe argumenty na rzecz efektywności eksplanacyjnej zintegrowanego modelu złożonych działań intencjonalnych przedstawione zostaną w ostatnim rozdziale pracy.

1.7 Plan pracy

Określenie wpływu stanów intencjonalnych na dobór zachowań wymaga stosownego dookreślenia pojęcia intencjonalności. Jak wiadomo, do współczesnej filozofii wprowadził je Franz Brentano (Brentano, 1999), nadal jednak jest ono przedmiotem licznych kontrowersji oraz sporów interpretacyjnych. W pracy wykorzystuję rozumienie tej kategorii zaproponowane przez Johna Searle'a. Zainspirowany teorią intencjonalności opracowaną przez amerykańskiego filozofa doprecyzowuję i włączam do zintegrowanego modelu kluczowe – z perspektywy struktury i funkcji działania intencjonalnego – typy stanów intencjonalnych decydujących o jego przebiegu (prior intencję, intencję w działaniu oraz przekonania).

Na obecnym etapie rozwoju kognitywistyki określenie związków między poszczególnymi składowymi działania intencjonalnego możliwe jest tylko w sposób

przybliżony. Wiedza na temat neuronalnej implementacji stanów intencjonalnych nie pozwala jednoznacznie stwierdzić, jaki jest wpływ, odpowiadających tym stanom, struktur neuronalnych na dobór programów motorycznych realizujących poszczególne zachowania. W tej sytuacji należy zadać pytanie: jak możliwa jest przyczynowość mentalna bez popadania w rozwiązanie dualistyczne, nieakceptowalne (gdyż nie daje się ono testować metodami empirycznymi) z perspektywy naukowej? Konstruktywnej odpowiedzi na tak postawione pytanie udzielił John Searle w eseju na temat intencjonalności zatytułowanym *Intentionality, an Essay in the Philosophy of Mind*. Prezentacja argumentacji filozofa z Berkeley jest zadaniem wieloetapowym, które wymaga m.in. wprowadzenia aparatu pojęciowego służącego do doprecyzowania definicji intencjonalności, wyróżnienia istotnych – z perspektywy działań – typów stanów intencjonalnych oraz reguł rządzących działaniami. Przedstawione zagadnienia są głównym przedmiotem dociekań w **rozdziale 2.,** zatytułowanym *Przyczynowy wpływ stanów intencjonalnych na wybór zachowań*. Efektem tych rozważań jest schemat działania intencjonalnego stanowiący ramę pojęciową oraz punkt odniesienia dla zagadnień poruszanych w kolejnych rozdziałach pracy.

Rozdział 3. poświęcony jest neurobiologicznemu mechanizmowi odpowiedzialnemu za organizowanie zachowań w uporządkowane sekwencje. Rdzeniem tego typu mechanizmu, zgodnie z hipotezą dopaminergicznego błędu predykcji nagrody (HDBPN), jest algorytm TDRL (*Temporal Difference Reinforcement Learning*), implementujący metodę uczenia się ze wzmacnianiem. W rozdziale tym omawiam zasadę działania oraz najważniejsze cechy algorytmu TDRL. Analizy modelu obliczeniowego uwzględniającego HDBPN prowadzą do określenia zakresu jego stosowalności w odniesieniu do złożonych działań intencjonalnych. Punktem wyjścia tych analiz jest hipoteza „super-mocy" Reada Montague'a, za pomocą której amerykański neuronaukowiec wyjaśnia specyficzne dla gatunku ludzkiego zachowania, naruszające powszechnie przyjmowaną zasadę dominacji instynktu przetrwania. Ważne jest, by w tym kontekście zastanowić się nad tym, w jakim stopniu interpretacja zaproponowana przez Montague, a dotycząca mechanizmu uczenia się ze wzmacnianiem, pozwala wyjaśnić podstawowe cechy działań intencjonalnych wraz z ich charakterystyczną dynamiką, czyli stopniowym przesuwaniem działań rutynowych w kierunku zdolności tła.

W **rozdziale 4.** przedstawiam i krytycznie analizuję najważniejsze wyniki badań realizowanych w psychologii intencji. Dane eksperymentalne, jak już wspomniałem, pozwalają wniknąć w strukturę zidentyfikowanej przez Searle'a intencji w działaniu oraz

w jej „fenomenalne otoczenie". Pewne rozstrzygnięcia posiadają bogatą ewidencję empiryczną, np. badacze zgadzają się co do korelacyjnego, a nie przyczynowego charakteru określonych składowych intencji. Status innych fenomenów, takich jak poczucie sprawstwa, budzi szereg wątpliwości i polemik (Bayne, 2006; Haggard, 2005; Wegner, 2002). Z perspektywy celu niniejszej pracy najważniejszą częścią tego rozdziału jest analiza funkcjonalnych aspektów tych stanów. Rezultaty przeprowadzonych analiz wykorzystuję w ostatnim rozdziale pracy.

W **rozdziale 5.** przedstawiam i szczegółowo dyskutuję zintegrowany model złożonych działań intencjonalnych. Prezentacja modelu odbywa się w dwóch etapach. Etap pierwszy poświęcony jest sformułowaniu najważniejszych wymagań funkcjonalnych wobec modelu. Ich podstawą są wyniki wcześniejszych ustaleń. Na etap drugi składają się charakterystyki poszczególnych podsystemów modelu oraz jego wybrane konkretyzacje, ukazujące wpływ poszczególnych składników modelu na jego cechy funkcjonalne.

2 Przyczynowy wpływ stanów intencjonalnych na wybór zachowania

Kiedy wykonujemy codzienne obowiązki, często umyka nam, jak złożona jest struktura naszych pozornie prostych zachowań. Nie zdajemy sobie sprawy, przygotowując posiłek, biegnąc do tramwaju, używając komputera lub prowadząc samochód, jak skomplikowany jest mechanizm kontrolujący ruchy naszego ciała i jak złożone są procesy umysłowe (percepcyjne, decyzyjne, kontrolne, wolicjonalne, itp.) leżące u podstaw wymienionych aktywności. Dopiero kiedy obserwujemy zachowania osób wykonujących nieznany nam zawód, np. maklerów giełdowych z Wall Street, zauważamy, jak trudno jest zrozumieć to, co widzimy: dziwne gesty, nerwowe okrzyki, wpatrywanie się w ekrany monitorów z kolumnami zmieniających się ciągów cyfr, robienie notatek, niezliczone rozmowy telefoniczne, itd. Wszystkie te działania będą dla nas niezrozumiałe, dopóki ktoś nam nie wyjaśni, skąd się one biorą i jaki jest ich sens, czyli jakie intencje, pragnienia i przekonania sprawiły, że maklerzy postępują w taki, a nie w inny sposób. Bez odpowiednich wyjaśnień, wskazujących określone procesy umysłowe jako źródło obserwowanych działań, trudno się pozbyć myśli, że patrzymy na „dziwne" zachowania, z którymi nie mieliśmy wcześniej do czynienia. W tego typu przypadkach ujawnia się istota działania intencjonalnego jako splotu (1) procesów umysłowych odnoszących się do otaczającego nas świata oraz (2) zachowań (ruchów ciała, wokalizacji, kontroli motoryczno-sensorycznej, itp.). Dopiero odwołanie się do szczególnego rodzaju powiązania mechanizmów umysłowych i cielesnych tłumaczy tę charakterystyczną dla osobników naszego gatunku zdolność do działań, które interpretujemy jako inteligentne.

W kognitywistyce toczy się obecnie spór o to, na ile – podczas analizy zachowań lub procesów poznawczych – obydwie formy aktywności powinny być rozpatrywane łącznie, a na ile można je traktować jako niezależne, wytworzone przez dwa odrębne, choć współpracujące ze sobą, podsystemy wykorzystujące specyficzne dla siebie mechanizmy i

reprezentacje[13] (Clark & Toribio, 1994). Z perspektywy obliczeniowej efektywniejsze jest podejście klasyczne, w którym się postuluje wyodrębnienie wyspecjalizowanych modułów, gdyż pozwala ono rozłożyć cały problem działania intencjonalnego na podproblemy składowe i opracować dla każdego z nich dedykowane rozwiązania, tj. opisać mechanizmy ich funkcjonowania oraz właściwe dla danego podsystemu typy reprezentacji. Z drugiej strony, jak twierdzi Terrence Sejnowski (P. S. Churchland & Sejnowski, 1992), systemy biologiczne nie dają się łatwo wtłoczyć w wysoce wyidealizowane konstrukcje teoretyczne wykorzystywane w podejściach obliczeniowych. Rozwiązaniem tego dylematu jest przyjęcie, że modele obliczeniowe są idealizacjami, które należy stopniowo „urealistyczniać", aby wyjaśnić działanie konkretnego systemu biologicznego. To urealistycznienie polega na konstruowaniu kolejnych, coraz bardziej złożonych modeli, które w coraz wyższym stopniu przybliżają struktury i funkcje organizmów biologicznych. Odnosi się to w szczególności do obliczeniowego modelowania działań intencjonalnych. Proste, najbardziej wyidealizowane, modele takich działań trafniej charakteryzują funkcjonowanie sztucznych systemów inteligentnych, niż żywych istot. Dopiero kolejne wersje takich modeli można odnieść do organizmów biologicznych, poczynając od mniej złożonych, a kończąc na reprezentantach *homo sapiens*. W niniejszej pracy przyjąłem takie właśnie wielostopniowe podejście, które polega na konstruowaniu – w trybie kolejnych przybliżeń – sekwencji modeli pozwalających zademonstrować zarówno wewnętrzną strukturę systemu odpowiedzialnego za konstrukcję i realizację działań intencjonalnych, jak i jego funkcjonalne możliwości. Prezentację i objaśnienie tych modeli zawiera ostatni rozdział niniejszej rozprawy. Proponowane w nim podejście różni się od większości dotychczasowych ujęć. Te ostatnie skupiają sią na ustalaniu relacji między zgrubnie scharakteryzowanymi „modułami": intencjonalnym, czyli umysłowym oraz zachowaniowym, czyli motorycznym. Koncentracja na tym, jak przezwyciężyć opozycję między tym, co umysłowe a tym, co cielesne sprawia, że zwolennicy takiego podejścia znacznie mniej uwagi poświęcają wewnętrznej strukturze każdego z modułów. Tymczasem, wniknięcie w to, jak w istocie zbudowany jest „moduł intencjonalnej kontroli" – pozwala dostrzec, że jest to kompleks wyspecjalizowanych podsystemów, które trudno

[13] W opinii Clarka i Toribio, krytyka podejścia reprezentacjonistycznego bierze się w dużej mierze ze zbyt wąskiego pojmowania terminu reprezentacja. „Our claim will be that the empirically driven anti-representationalist vastly overstates her case. Such overstatement is rooted, we suggest, in an unwarranted conflation of the fully general notion of representation with the vastly more restrictive notions of explicit representation and/or of representations bearing intuitive, familiar contents." (Clark & Toribio, 1994, s. 2). Odpowiednio rozszerzone pojęcie reprezentacji obejmujące zarówno ujęcie koneksjonistyczne jak i symboliczne (Haugeland, 1998) pozwala rozwiać zgłaszane przez antyreprezentacjonistów wątpliwości.

wywieść z takich jednostek (przeżyć) umysłowych, jak przekonania, zamiary, sądy wartościujące, itp. To samo odnosi się do modułu zachowaniowego, który także złożony jest z podsystemów. Ich zadaniem jest wypracowanie – w trybie uczenia się – finalnego zachowania, które w przeciwieństwie do mimowolnych ruchów ciała ma zawsze charakter celowy.

Nie znaczy to bynajmniej, że dotychczasowe, w mojej opinii nadmiernie uproszczone, analizy dotyczące działań intencjonalnych są całkowicie chybione. John Searle należy do tych autorów, którzy zdają sobie sprawę ze złożonej natury działania intencjonalnego. Koncepcję amerykańskiego filozofa omówię dokładniej, gdyż zawiera ona szereg cennych spostrzeżeń i propozycji. Nawiążę do nich w ostatnim rozdziale pracy, przedstawiając własne, modelowe podejście do charakterystyki złożonego działania intencjonalnego.

John Searle przedstawił w 1983 roku oryginalną propozycję dotyczącą działań rozumianych jako splot tego, co umysłowe i tego, co motoryczne. Opracowany przez niego aparat pojęciowy oferuje spójny, przejrzysty i heurystycznie płodny obraz zarówno prostych, jak i złożonych działań intencjonalnych. Ludzkie zachowanie, jak zauważa amerykański filozof, w zasadzie nigdy nie daje się zredukować tylko do tego, co czysto motoryczne. Taki sam – z fizycznego punktu widzenia – ruch ręki służyć może do wykonywania zupełnie różnych działań, choćby takich jak: udział w głosowaniu, gest wyrażający radość, przywitanie, podziękowanie, zgłoszenie się do odpowiedzi. Wszystkie wymienione przypadki różni głównie kontekst kulturowy oraz rodzaje procesów umysłowych przypisywanych podmiotom tych działań. Ponadto, zdecydowana większość działań intencjonalnych to najczęściej sekwencje zachowań, złożone z wielu prostych, dostrajanych w procesie uczenia się, ruchów ciała. Ruchy te są składnikami szerszego planu, który, zanim zostanie motorycznie zrealizowany, wymaga na ogół opracowania pojęciowego. Złożone działania nie byłyby możliwe, gdyby nie poprzedzały ich odpowiednie operacje na reprezentacjach mentalnych obiektów z otoczenia. Dotyczy to zwłaszcza współczesnego środowiska kulturowego nasyconego mnogością artefaktów, w szczególności tych, które są wysoce zaawansowane technicznie. Analiza tak złożonego zjawiska, jakim jest działanie intencjonalne oraz wydobycie jego najistotniejszych składników, wymaga odpowiednio pojemnego aparatu pojęciowego. Jak już wcześniej wspomniałem, pomocna w przygotowaniu takiej analizy jest teoria intencjonalności Johna Searle'a. Choć została ona skonstruowana na potrzeby polemiki z dominującymi w filozofii

umysłu stanowiskami: materializmem i dualizmem[14], to opracowana w niej aparatura pojęciowa nadaje się do systematycznej rekonstrukcji najważniejszych cech działań intencjonalnych. Z perspektywy zasadniczego celu niniejszej rozprawy szczególnie ważne jest to, że amerykański filozof opowiada się za uwzględnieniem wiedzy empirycznej o działaniu układu nerwowego oraz biologicznych podstawach zachowań ludzkich. Barry Smith twierdzi, że dane naukowe są dla Searle'a swoistym „zabezpieczeniem", które minimalizuje ryzyko popadnięcia w intelektualne nonsensy (Smith, 2003, s. 1).

Niniejszy rozdział poświęcony jest analizie relacji między określonymi typami umysłowych stanów intencjonalnych a zachowaniami. W szczególności, interesować mnie będzie odpowiedź na pytanie: czy można zasadnie postulować, by tego typu relacje miały status związku przyczynowego? Podstawą prowadzonych rozważań będzie filozoficzna teoria intencjonalności Johna Searle'a, która umożliwi przygotowanie odpowiedzi m.in. na powyższe pytanie.

Odniesienie się do tak sformułowanego problemu wymaga: doprecyzowania pojęcia „intencjonalności", przeanalizowania stanów mózgu-umysłu będących nośnikami intencjonalności oraz określenia relacji występujących między tymi stanami. Dysponując tego typu aparatem pojęciowym można przejść do zbadania, a następnie wyjaśnienia związków pomiędzy wybranymi typami stanów intencjonalnych a zachowaniami. Przyjmuję, za Searlem, że najważniejsze tego typu stany to: prior intencja, intencja w działaniu oraz przekonania. W charakterystyczny dla siebie sposób determinują one dobór odpowiednich zachowań. Analiza zachodzących między nimi relacji pozwala na nowo spojrzeć na powiązane z nimi związki przyczynowo-skutkowe. Mogłoby się wydawać, że od czasów Hume'a panuje zgoda odnośnie sposobu funkcjonowania tego typów związków, jednak, jak twierdzi Searle, ujęcie autora *Traktatu o naturze ludzkiej* nie jest trafne i widać to szczególnie wyraźnie, gdy odniesie się je do działań intencjonalnych (Searle, 1983, s.

[14] W filozofii umysłu dominują obecnie koncepcje materialistyczne (patrz: różne wersje teorii identyczności [U.T. Place, E. G. Boring, J.J.C. Smart], funkcjonalizm [H. Putnam, J. Fodor], eliminatywizm [P. i P. Churchland], monizm anomalny [D. Davidson]). Wśród prominentnych filozofów można również spotkać zwolenników dualizmu własności (D. Chalmers), a nawet nowych mysterianistów (C. McGinn). Searle w dużym stopniu utożsamia się ze stanowiskiem materialistycznym, choć krytykuje niektóre z jego wersji. Oprócz szczegółowych zarzutów wobec poszczególnych stanowisk (np. krytyka silnej AI oraz podejścia komputacyjnego, krytyka eksternalizmu semantycznego oraz tzw. trudnego problemu świadomości), Searle często zarzuca zwolennikom materializmu, że w swych analizach pomijają oczywisty fakt, mianowicie, że mózg-umysł to twór biologiczny, produkt procesu ewolucji. Kiedy próbuje się ten fakt zignorować, łatwo popaść w konstrukcje prowadzące do jałowych sporów i problemów. Autor *Umysłu na nowo odkrytego*, aby odróżnić własną koncepcję od innych dominujących w filozofii umysłu, ukuł dla swojego stanowiska określenie „naturalizm biologiczny", za pomocą którego pragnął zwrócić uwagę na konteksty, w których prowadzi rozważania.

183). W postulowanej przez Hume'a redukcji związku przyczynowo-skutkowego do wytworu wyobraźni (czyli określonej operacji umysłowej na dwóch stycznych i stale współwystępujących zdarzeniach) nie można uchwycić specyfiki naszego codziennego doświadczenia związanego choćby z „fizycznymi" doznaniami podczas kolizji z jakimś obiektem (np. utraty równowagi oraz doznania bólu spowodowanego faulem gracza drużyny przeciwnej podczas meczu piłki nożnej). Ponadto, w ujęciu tym nie uwzględniono silnego poczucia, że zamiary człowieka w istotny sposób decydują o zmianach wywoływanych przez niego w świecie (np. jeśli postanowię zamknąć okno w pokoju, to zostanie ono zamknięte, o ile coś lub ktoś mi nie przeszkodzi). Omówienie głównych pojęć Searle'owskiej teorii intencjonalności oraz rekonstrukcja jego poglądów na strukturę i dynamikę działania intencjonalnego będzie „pierwszym przybliżeniem" mechanizmów odpowiedzialnych za jego konstrukcję i realizację, czyli wstępnym, najbardziej zgrubnym jego modelem.

2.1 Struktura intencjonalności

Stan intencjonalny

Przyjmuję za Johnem Searlem, iż „intencjonalność to własność wielu stanów psychicznych, polegająca na przedstawianiu czegoś, skierowaniu na coś czy też na byciu o czymś – «coś» w tym przypadku to albo przedmioty, albo stany rzeczy znajdujące się poza nimi" (Searle, 1983, s. 2). Przykładami tego typu stanów są:

- percepcje, np. „widzę biurko", „słyszę muzykę",
- odczucia odnoszące się do stopnia zaspokojenia podstawowych potrzeb organizmu, np. „jestem głodny",
- określone emocje, np. lęk wywołany widokiem węża,
- przekonania, np. „Ziemia jest jedną z planet Układu Słonecznego",
- pragnienia, np. „pragnę zwyciężyć w kampanii prezydenckiej",
- zamiary, np. „zamierzam zabić mojego prześladowcę".

Potraktowanie intencjonalności jako szczególnego rodzaju relacji między stanem organizmu (przede wszystkim stanem mózgu) a faktycznym albo projektowanym stanem otoczenia odbiega od rozumienia jej zgodnie z tradycją brentanowsko-husserlowską. Jednak za sprawą takich m.in. badaczy jak Fred Dretske (2004), Ruth Garrett Millikan

(2009) czy John Searle (1983) upowszechniła się naturalistyczna interpretacja tej relacji. Krystyna Bielecka (2019) w książce *Błądzę, więc myślę. Co to jest błędna reprezentacja?* dokonała obszernego przeglądu znaturalizowanych ujęć reprezentacji umysłowych, a w szczególności tych, które mają charakter intencjonalny.

Można powiedzieć, używając terminu z obliczeniowej teorii umysłu (*the Computational Theory of Mind*), że stany intencjonalne są **reprezentacjami umysłowymi** odnoszącymi ich posiadacza do świata w jego najróżniejszych przejawach. Pojawia się zatem pytanie: jak tego typu funkcja realizowana jest przez układ nerwowy?

Treść oraz modus psychologiczny

Zgodnie z ujęciem Searle'a – funkcja reprezentowania realizowana przez stany intencjonalne opiera się na dwóch elementach: (a) treści reprezentacyjnej 'p' oraz (b) modusie psychologicznym 'S'. Pełny stan intencjonalny jest zatem złożeniem wymienionych elementów, czyli S(p).

Treść reprezentacyjna umysłowego stanu intencjonalnego odsyła do procesu albo stanu świata (otoczenia). Taki proces albo stan jest zatem odniesieniem przedmiotowym stanu intencjonalnego. Warto dodać, że treści dane są nam zawsze poprzez określone wyglądy (mają charakter aspektowy) i mogą się odnosić zarówno do stanów świata (reprezentowanych na poziomie języka w formie zdań), jak i do konkretnych przedmiotów (reprezentowanych przez nazwy lub imiona własne). Modus psychologiczny określa typ stanu umysłowego (spostrzeżenie, wyobrażenie, pragnienie, zamiar, przekonanie, itp.), za pośrednictwem którego dostępna jest dana treść reprezentacyjna.

Ważne jest, że ta sama treść w połączeniu z różnymi modusami tworzy odmienne stany intencjonalne. Przykładowo, jeśli za treść przyjąć „padający deszcz" (p), to treść ta może być nam dana w formie pragnienia: „pragnę, by zaczął padać deszcz" S1(p), w formie przekonania: „jestem przekonany, że zacznie padać deszcz" S2(p), obawy: „obawiam się, że będzie padać deszcz" S3(p), nadziei: „mam nadzieję, że będzie padać deszcz" S4(p), itd. Zauważmy, że tym, co się zmienia w powyższych przykładach, są modusy psychologiczne, natomiast treść reprezentacyjna pozostaje taka sama. W tym kontekście nasuwa się pytanie: czym w istocie różnią się poszczególne „modusy psychologiczne" S1, S2, …, Sn?

Nakierowanie na zgodność oraz warunki spełniania

Należy zauważyć, poszukując różnic między modusami z powyższego przykładu (S1, S2, S3 oraz S4), że niemal dla każdego stanu intencjonalnego charakterystyczne jest swoiste „nakierowanie na zgodność" (Searle, 2010b, s. 169). Stany intencjonalne (rozumiane tu w duchu Searle'a jako układy złożone z modusu i treści), są – mówiąc obrazowo – „zainteresowane", by to, co w nich prezentowane – było **zgodne** z tymi składnikami lub aspektami świata, do których się odnoszą. Czasami to zainteresowanie polega na „dostosowaniu" odniesienia przedmiotowego wskazanego za pomocą treści reprezentacyjnej do faktycznego stanu otoczenia (są to m.in. przypadki percepcyjnych stanów intencjonalnych, takich jak: „widzę", „słyszę", „czuję" oraz stanów przekonaniowych takich jak: „wiem", „przypuszczam"), a czasami na potraktowaniu treści jako projektu wprowadzenia zmian (oczekiwanych, postulowanych, wyobrażonych, itp.) w aktualnym otoczeniu. W tym drugim przypadku aktualny stan otoczenia jest „dostosowywany" do wskazanego w treści reprezentacyjnej stanu świata (np. „z utęsknieniem czekam aż zacznie/przestanie padać deszcz"). Przekonania, hipotezy, spostrzeżenia zmysłowe, przypomnienia, itp. są przykładami stanów, których zadaniem jest stworzenie prawdziwej reprezentacji stosownych składników lub aspektów aktualnego lub minionego świata. Jeśli przypomnienie ma się do czegoś odnosić, to fakt z przeszłości musi odpowiadać treści tego przypomnienia. Podobnie, przekonanie oraz hipoteza odnoszą się do czegoś, o ile istnieją odpowiadające im stany świata. Mówimy, że stan intencjonalny jest prawdziwy, gdy składająca się na niego treść jest zgodna z pewnym faktem w świecie. W przeciwnym przypadku stan intencjonalny staje się fałszywy, gdyż nie ma zgodności między tymże stanem a rzeczywistością. Z kolei pragnienia, plany, zamiary będą się do czegoś odnosić, gdy świat zostanie odpowiednio zmodyfikowany – świat musi się zmienić, by to, co stan intencjonalny projektuje – okazało się rzeczywiste. Pragnienie napicia się wody zostanie zaspokojone, jeśli spragniony faktycznie napije się wody.

Abstrakcyjnie rzecz ujmując, nakierowanie na zgodność dzieli przestrzeń stanów intencjonalnych na trzy typy: (1) stany, w których nakierowanie na zgodność zachodzi w kierunku: umysł→świat (np. przekonania), (2) stany, w których zgodność zachodzi w kierunku: świat→umysł (np. pragnienia) oraz (3) stany o zerowym nakierowaniu na zgodność np. stan odnoszący się do stwierdzenia: jest mi przykro, że Cię zraniłem (ten typ nakierowania zostanie pominięty w niniejszej pracy ze względu na jego mniejsze znaczenie w perspektywie prowadzonych analiz).

Warto od razu zaznaczyć, że obydwa typy nakierowania na zgodność mają uzasadnienie ewolucyjne. Trudno bowiem wyobrazić sobie, by organizmy wytwarzające zasadniczo nieadekwatne stany intencjonalne mogły przetrwać oraz osiągnąć sukces reprodukcyjny. W tym kontekście, jak twierdzi Searle, szczególnie istotne są stany o niezerowym nakierowaniu na zgodność.

> *W każdym wypadku, gdy mamy do czynienia ze stanem intencjonalnym, który ma niezerowe nakierowanie na zgodność, zgodność ta może zaistnieć albo nie – przekonanie okaże się prawdziwe, pragnienie zaspokojone, a zamiar zrealizowany albo też nie (zależnie od konkretnego przypadku). […] wszystkie stany intencjonalne o niezerowym nakierowaniu na zgodność spełniane są na mocy określonych warunków, które nazywam warunkami spełniania. Same zaś stany intencjonalne należy rozumieć jako reprezentacje warunków, pod jakimi zostają spełnione* (Searle, 2010, s. 171).

Zgodnie z tym ujęciem, stan intencjonalny S(p) prezentuje *de facto* odpowiednio zorganizowane (modi-S) warunki spełniania (treść-p). Jeśli, przykładowo, widzę szklankę, to stan intencjonalny mego umysłu będzie spełniony, o ile w polu mojego widzenia w tej właśnie chwili znajdować się będzie szklanka, a spostrzegane przeze mnie cechy tego obiektu (takie jak: kształt, wielkość i barwa) będą wiernym odwzorowaniem cech stojącej przede mną rzeczy. Podobnie będzie w przypadku pragnienia „chcę schronić się przed deszczem" – tego typu stan intencjonalny zostanie spełniony, gdy przykładowo wyciągnę parasol lub schowam się pod drzewem, innymi słowy, gdy w odpowiedni sposób wpłynę na otaczającą mnie rzeczywistość poprzez ochronę przed zmoczeniem mojej odzieży i ciała.

Tryby reprezentowania

Funkcja reprezentowania realizowana przez poszczególne stany intencjonalne może być realizowana – w zależności od organizacji warunków spełniania – w trybie bezpośredniej lub pośredniej prezentacji.

Z trybu bezpośredniego „korzystają" stany percepcyjne oraz stany odpowiedzialne za dobór działań, tj. prior intencja oraz intencja w działaniu (doprecyzowanie znaczenia tych pojęć nastąpi w sekcji *Schemat pełnego działania intencjonalnego*). Jak twierdzi Searle, szczególną właściwość ma układ warunków spełniania tego typu stanów, mianowicie część z nich wchodzi w relacje przyczynowo-skutkowe ze stanami świata. Bardzo dobrze

ilustruje to przykład z obszaru percepcji wzrokowej. Kiedy widzę przed sobą samochód, to jest to skutek tego, że odpowiednio odbite fale świetlne wywołują (działanie przyczynowe) w umyśle określone warunki spełniania, które poprzez treść odnoszą się do prezentowanego samochodu. Analogicznie funkcjonuje zamiar (*intention*) podjęcia działania, np. zamiar podniesienia ręki. Tego typu stan zawiera dwie powiązane ze sobą grupy warunków spełniania. Pierwsza odnosi się do treści zamiaru, którą w tym przypadku jest odczuwanie działania (*experience of acting*) polegające np. na dążeniu do podnoszenia ręki, druga natomiast do samego zachowania, którego efektem będzie to, że w określonym momencie ręka zostanie podniesiona. Zdaniem Searle'a, pomiędzy pierwszą a drugą grupą istnieje relacja przyczynowo-skutkowa[15]. Dążenie do podniesienia ręki jest przyczyną tego, że zostaje ona podniesiona.

Drugi tryb – pośrednie przedstawianie – różni się od bezpośredniej prezentacji tym, że warunki spełniania tego typu stanów nie muszą prowadzić do wystąpienia określonych związków przyczynowo-skutkowych. Ich cechą jest odłączalność[16] od tego, do czego się odnoszą. Przekonania, pragnienia, wyobrażenia, itd. funkcjonują na tej właśnie zasadzie. Jeśli, na przykład, odczuwam chęć, by napić się wody, to w treści tego stanu znajdą się warunki wskazujące na sposób jego zaspokojenia (np. wypicie wody ze szklanki), jednak nie będą one miały tej samej mocy przyczynowej, co zamiar jej wypicia. Innymi słowy, będąca skutkiem pragnienia chęć napicia się wody w odróżnieniu od zamiaru napicia się wody[17] może być zaspokojona (spełniona) na wiele różnych sposobów, liczy się bowiem jedynie końcowy efekt, a nie sposób jego osiągnięcia. Natomiast możemy jednoznacznie orzec o zamiarze, który – pomimo pojawienia się nadarzającej się sposobności – nie został zrealizowany, że był to zamiar pozorny, tzn. taki, który albo nie był autentyczny, albo został porzucony. W tym przypadku realizacja stanowi konstytutywny element zamiaru. Dobrze ilustruje to przykład osoby prowadzącej radykalną głodówkę, a więc odmawiającej przyjmowania płynów. Osoba taka z pewnością ma chęć napicia się wody, jednak świadomie rezygnuje z zamiaru jej wypicia.

[15] "In each case there is a self-referential Intentional state or event, and the form of the self-reference (in the case of action) is that it is part of the content of the Intentional state or event that its conditions of satisfaction (in the sense of requirement) require that it cause the rest of its conditions of satisfaction (in the sense of thing required) […]. When I raised my arm, I directly experienced the causing: I did not observe two events, the experience of acting and the movement of the arm, rather part to the Intentional content of the experience of acting was that that very experience was making my arm go up. […] I can directly experience the relation of one thing making another thing happens." (J. R. Searle, 1983, s. 122).

[16] Searle używa w tym kontekście terminu *detachable*.

[17] Jak pokażę poniżej, Searle dookreśla to pojęcie, odróżniając prior intencję od intencji w działaniu.

Sieć stanów intencjonalnych

Zdefiniowanie stanu intencjonalnego jako S(p) pozwala wyróżnić w nim element treściowy i psychologiczny, określić nakierowanie na zgodność (świat→umysł, umysł→świat) oraz przypisać go do jednej z wielu kategorii (przekonania, pragnienia, lęki, itd.). Jednak stany intencjonalne, zdaniem Searle'a, nie są wyizolowanymi jednostkami. Każdy stan umysłowy/intencjonalny jest tylko jednym ze składników złożonej sieci takich stanów i tylko ze względu na tę sieć można rozpatrywać warunki jego spełniania.

> *Moglibyśmy się nawet pokusić o twierdzenie, że funkcjonowanie dowolnego stanu intencjonalnego (które rozumiemy tu jako determinowanie przezeń swoich warunków spełniania) możliwe jest tylko względem sieci intencjonalności, której jest on częścią*. (Searle, 2010b, s. 174).

Taki „sieciowy" charakter stanów intencjonalnych jest wynikiem szeregu relacji, w które wchodzą ze sobą poszczególne stany za pośrednictwem swoich warunków spełniania. Searle rozważa następującą prior intencję jako przykład obrazujący wzajemne zależności między stanami intencjonalnymi: „zamierzam wystartować w wyścigu o fotel prezydenta Stanów Zjednoczonych". Aby taki zamiar mógł pojawić się w umyśle danej osoby, musi ona posiadać szereg dodatkowych stanów, które „umożliwią" jego zaistnienie. Do najbardziej prawdopodobnych należą:

1. Przekonanie S_{prz}(p1) – „Stany Zjednoczone Ameryki Północnej to republika".
2. Przekonanie S_{prz}(p2) – „W USA odbywają się cyklicznie wybory prezydenta".
3. Przekonanie S_{prz}(p3) – „W USA dominują dwie partie, których kandydaci walczą o prezydenturę".
4. Pragnienie S_{prag}(p4) – „Pragnę pozyskać nominację mojej partii".
5. Itd. (Searle, 1983, s. 141).

Powyższa lista stanów intencjonalnych to tylko przykład. W konkretnym przypadku może ona przybrać inną postać, z pewnością jednak tego typu zamiar nie może pojawić się samoistnie jako samodzielny stan niepowiązany ze współdeterminującą go siecią innych stanów intencjonalnych. Searle zadaje retoryczne pytanie, aby pokazać to jeszcze wyraźniej: czy zamiar startu w wyborach prezydenckich mógłby pojawić się w umyśle człowieka z epoki kamienia łupanego?

Pojęcie sieci intencjonalnej może sugerować, że współtworzące ją składniki (stany intencjonalne) mają charakter dobrze wyodrębnionych jednostek, które łatwo jest zidentyfikować i policzyć. W rzeczywistości jest to znacznie bardziej skomplikowane. Granice pomiędzy poszczególnymi stanami są płynne i przez to trudno je jednoznacznie wyodrębnić. Szczególnie wyraźne jest to w przypadku percepcji, która ma charakter ciągły, podobnie jest w przypadku przekonań i pragnień, które są nam dostępne w formie zdań języka naturalnego. Searle twierdzi, że efekt zależności i wielorakich relacji dotyczy również pojęć, dlatego jest on zwolennikiem tzw. zasady holizmu znaczeniowego. Złożoność sieci stanów intencjonalnych oraz kłopoty z wyodrębnieniem jej składników sprawiają, że nie da się, nawet z dużym przybliżeniem, ocenić – z iloma stanami intencjonalnymi mamy w jej obrębie do czynienia. Dodatkową przeszkodą jest fakt, iż wiele z naszych przekonań, pragnień, obaw czy lęków nie jest nam dane w sposób jawny, tzn. z istnienia wielu z nich nie zdajemy sobie sprawy. Należą one do nieświadomych stanów umysłowych, choć, jak twierdzi Searle, potencjalnie każdy z nich może zostać uświadomiony[18]. Gdyby nawet przyjąć, że udałoby się przezwyciężyć wymienione trudności w zidentyfikowaniu poszczególnych stanów intencjonalnych, to pojawia się jeszcze jedna przeszkoda w realizacji tego zadania, znacznie bardziej zasadnicza, niż dotychczas wymienione. Przeszkodą tą jest tło (*background*), czyli zbiór przedintencjonalnych dyspozycji, umiejętności, nastawień, zdolności, założeń, praktyk i nawyków, które współkonstytuują sieć stanów intencjonalnych.

Tło

> *Tło jest «przedintencjonalne» w tym sensie, że choć nie jest ono formą czy formami intencjonalności, to jest warunkiem wstępnym lub zbiorem warunków wstępnych intencjonalności* (Searle, 1983, s. 143).

Przykładami potwierdzającymi istnienie tła mogą być nasze oczekiwania wobec różnego rodzaju narzędzi czy przedmiotów codziennego użytku. Próbując podnieść duży kufel do piwa wykonany z bardzo lekkiego tworzywa – będziemy z pewnością zdziwieni i zaskoczeni jego niezgodną z oczekiwaniem niewielką wagą. Tego typu oczekiwania nie należy jednak rozumieć jako przekonania, że kufle do piwa są z reguły ciężkie. Jest to

[18] Nie istnieją, zdaniem Searle'a, stany intencjonalne, które potencjalnie nie mogłyby zostać uświadomione. Filozof ten jest przeciwnikiem tzw. głębokiej nieświadomości, która zawierałaby np. reguły przetwarzania stanów intencjonalnych, do których nigdy nie mielibyśmy świadomego dostępu.

raczej efekt szeregu wcześniejszych doświadczeń z obiektami o zbliżonej wielkości i kształcie. Obiekty te wcale nie musiały być kuflami do piwa. Tego rodzaju oczekiwanie nie jest typowym stanem intencjonalnym, funkcjonuje ono raczej jako niejawne „założenie" dotyczące pewnego fragmentu rzeczywistości. Gdy takie założenie zostanie w jakiś sposób „podważone" (np. kufel okaże się bardzo lekki), naszą reakcją będzie zdziwienie. Wówczas uświadomimy sobie, że „niejawny" stosunek do tego typu przedmiotów towarzyszył nam od zawsze[19]. Składniki tła, zdaniem Searle'a, nie są standardowymi reprezentacjami. Mogą one w określonych sytuacjach zostać „dostrzeżone" i wyrażone, jednak na tle jawnych przekonań wyróżniać je będzie szczególnego rodzaju „sztuczność i dziwność". Na przykład, trudno nie uznać za „dziwne" następującego twierdzenia: „W wyborach prezydenckich głosują osoby w pełni przytomne". Nie dość, że raczej nie przyszłoby nam do głowy sformułowanie takiego sądu, to na dodatek, nie da się go powiązać z naszymi typowymi przekonaniami. Skoro dyspozycje tła nie są standardowo włączone w sieć przekonań intencjonalnych, to pojawia się pytanie: w jaki sposób funkcjonują one w naszym umyśle oraz – jak z nich korzystamy? Zdaniem Searle'a, są one dostępne w formie dwojakiego rodzaju „wiedzy-jak" (*know-how*): (1) wiedzy, jakie są rzeczy (*how things are*) oraz wiedzy, jak coś zrobić (*how to do things*) (należy od razu zaznaczyć, że – zdaniem autora *Intentionality, an essay in the philosophy of mind* – wiedza-jak, w odróżnieniu od wiedzy-że, nie ma charakteru reprezentacyjnego).

Najprostszy zamiar (intencja), by napić się zimnego piwa, które znajduje się w lodówce, zaskakuje liczbą umiejętności opartych na wiedzy jak coś zrobić oraz jakie są rzeczy, które musi posiadać podmiot, by móc go zrealizować. W takim przypadku należy: potrafić i móc wstać z krzesła, zlokalizować lodówkę i podejść do niej, otworzyć drzwiczki, wyjąć butelkę, zamknąć lodówkę, otworzyć butelkę, napić się z butelki. Na ogół aktywacja tego typu umiejętności wymaga zaangażowania odpowiednich stanów intencjonalnych. W przytoczonym przykładzie będzie to uświadomienie sobie pragnienia oraz spostrzeżenie drzwi od lodówki, przy czym samo rozpoznanie danej powierzchni jako drzwi lodówki nie jest stanem intencjonalnym, a typową zdolnością tła. Jeśli przyjrzeć się wymienionym dyspozycjom pod kątem źródła ich pochodzenia, to łatwo dostrzec, że są to dyspozycje dwojakie: (1) biologiczne i (2) kulturowe. Chodzenie, jedzenie, spostrzeganie, rozpoznawanie, przedintencjonalne założenie dotyczące trwałości rzeczy, niezależności

[19] Efekt zdziwienia w tego typu przypadkach wskazuje, zdaniem Searle'a, że z poszczególnymi składnikami tła związane są również swoiste warunki spełniania.

obiektów oraz innych osób lub zwierząt (Searle, 1983, s. 144) – to przykłady dyspozycji wrodzonych, rozwijanych w toku rozwoju ontogenetycznego. Posługiwanie się narzędziami z kolei (m.in. otwieranie drzwi od lodówki, picie z butelki, posługiwanie się piłą, młotkiem, łukiem) – to przykład umiejętności nabywanej w zaprojektowanym w danej kulturze procesie uczenia się.

Kiedy rozpoczynamy naukę, to zazwyczaj towarzyszą nam proste wskazówki lub instrukcje od bliskich osób lub nauczycieli. W przypadku nauki jazdy na nartach mogą one przybrać postać następujących zaleceń: przenieś środek ciężkości ciała na odpowiednią stronę podczas skrętu; zegnij kolana; unikaj jazdy na wprost. Nieodłącznymi składnikami tych zaleceń są demonstracje właściwych zachowań. Instruktor oczekuje, aby uczeń naśladował jego ruchy, a wypowiedzi językowe wskazują, na co należy zwrócić szczególną uwagę. Im stajemy się bieglejsi w realizowaniu konkretnych czynności, tym mniej uwagi musimy poświęcać poszczególnym zaleceniom. W pewnym momencie poziom rozwoju danej umiejętność jest tak wysoki, że pojawia się efekt nieświadomej kompetencji: zamiast koncentrować się na ugięciu kolan czy na wykonaniu skrętu, po prostu jedziemy slalomem i niemal bezwiednie dostosowujemy ułożenie ciała do zmieniających się warunków zjazdowych. Biegłe opanowanie umiejętności sprawia, że świadoma kontrola działań, zamiast pomagać w realizacji celu, zaczyna przeszkadzać. Często w takich sytuacjach pojawiają się rady zupełnie inne niż poprzednio. Zamiast objaśnień i reguł słyszymy: „jedź, nie myśl”. Na ogół tego typu zaawansowaną umiejętność interpretuje się jako głęboką internalizację przekazanych nam w instrukcjach reprezentacji. Tymczasem, jak twierdzi Searle, przesunięcie danej umiejętności na poziom tła oznacza nadanie jej nowej jakości i niejako odcięcie jej od początkowych instrukcji. W rozdziale 4. doprecyzuję myśl Searle’a i pokażę, że efekt ten można traktować jako skutek długotrwałego procesu uczenia się ze wzmacnianiem.

Relacja: tło - stany intencjonalne

Searle twierdzi, próbując dookreślić relację między tłem a stanami intencjonalnymi, że:

> *[…] tło dostarcza zbioru warunków umożliwiających funkcjonowanie poszczególnych form intencjonalności. […] Tło określa warunki konieczne, ale niewystarczające do zrozumienia przekonań, pragnień, zamiarów, etc., i w tym sensie umożliwia ich istnienie, ale ich nie warunkuje.* (Searle, 1983, s. 157, 158).

Na podstawie powyższych stwierdzeń wnosić można, że treść stanu intencjonalnego – rozumiana jako zbiór warunków spełniania – zależy zarówno od innych stanów intencjonalnych (aspekt sieciowy stanu), jak i od nabytych wcześniej przedintencjonalnych dyspozycji tła. Relację między tłem a siecią stanów intencjonalnych zobrazować można za pomocą następującego schematu:

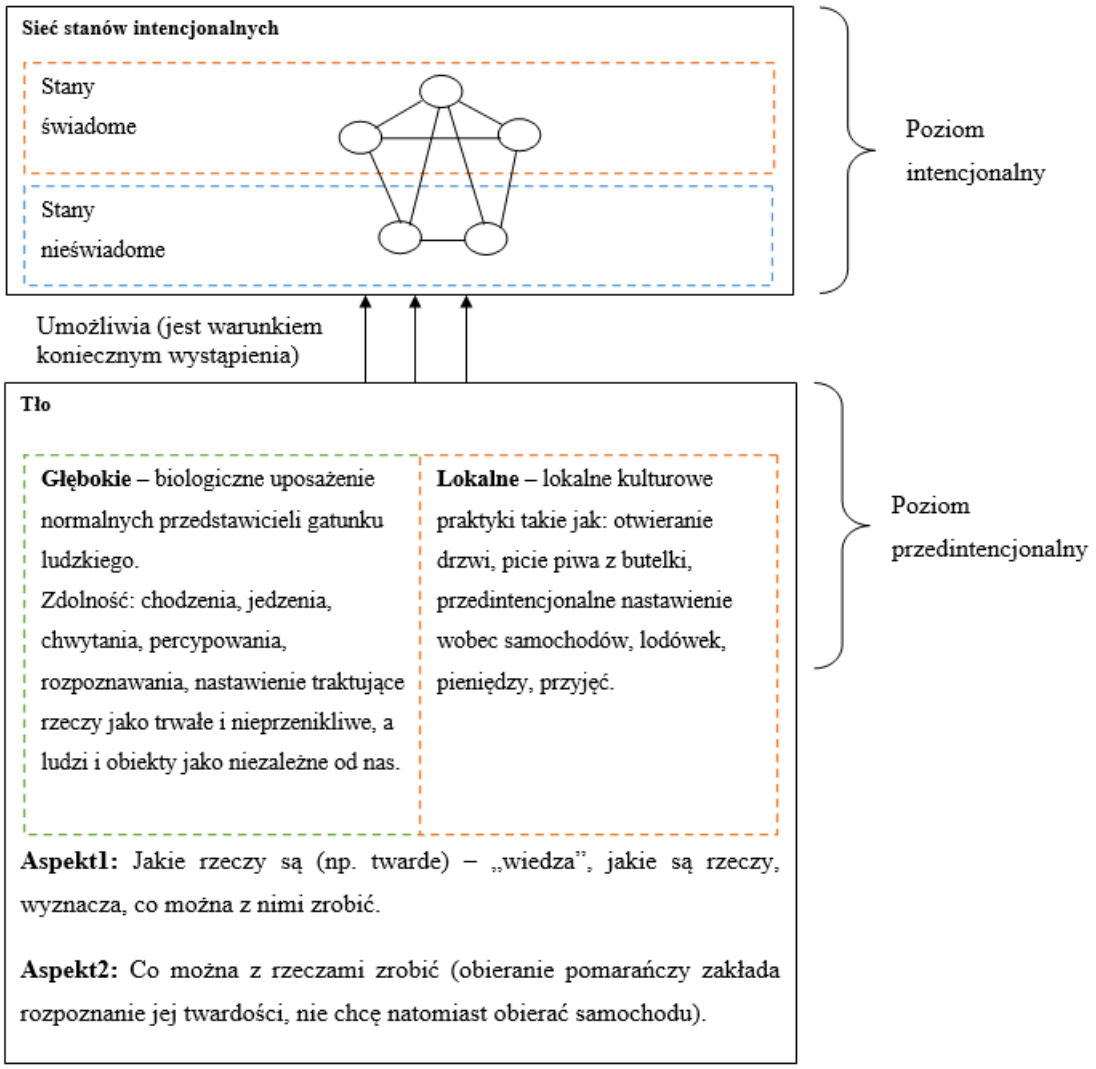


Diagram 2. Relacja między siecią stanów intencjonalnych a dyspozycjami tła.

Przedintencjonalny charakter tła sprawia, że jest ono trudno dostępne dla analizy pojęciowej. W szczególności, polega to na tym, że każda próba opisu powoduje „uintencjonalnienie" tych elementów tła, które próbuje się objaśnić lub doprecyzować. W efekcie, składnik tła zostaje ujęty jako wyłączony z niego akt intencjonalny. Kiedy tło warunkuje nasze zachowania, twierdzi Searle, wówczas „po prostu działamy", choć równocześnie błędem byłoby klasyfikowanie tego typu działań jako czysto mechanicznych. W tak zarysowanym kontekście powstaje pytanie: jak funkcjonują

działania intencjonalne, w których dochodzi do splotu dyspozycji tła oraz sieci stanów intencjonalnych?

2.2 Schemat pełnego działania intencjonalnego

Nawet pobieżna analiza działań dowolnych (*voluntary actions*) pokazuje, że występują one w wielu odmianach. Badacze wyróżniają działania zamierzone, akcydentalne, przemyślane, nieplanowane, spontaniczne, rutynowe, itp. Sprowadzenie ich wszystkich do jednego wzorca odsłania, co prawda, ich wspólną naturę, ale zarazem pozbawia każdą z tych form właściwej jej specyfiki. A to właśnie konkretna specyfika i mnogość typów działań intencjonalnych czynią umysł „kompleksem narzędzi" zdolnych nie tylko do reprezentowania aktualnych lub minionych stanów otoczenia, ale także do tworzenia reprezentacji stanów przyszłych i to takich, których realizacja wymaga zaplanowania czynności niezbędnych do jego zaistnienia. Często stany intencjonalne, które towarzyszą tego typu działaniom, różnią się między sobą w istotny sposób (mogą np. być spontaniczne, nieplanowane, planowane, itp.). W niniejszej pracy skupię się na modelowaniu szeroko pojętych działań zaplanowanych.

Searle rozpoczyna analizę działań intencjonalnych od objaśnienia różnicy między stanami intencjonalnymi powiązanymi przyczynowo z działaniami a stanami intencjonalnymi, które nie podlegają takim powiązaniom (percepcje, wyobrażenia, przypomnienia, itp.). Działania muszą być wykonane w prawidłowy sposób (*in the right way*), jak stwierdza amerykański filozof, tzn. zamiar musi pojawić się przed zachowaniem i to w ściśle określonym momencie. W przeciwnym przypadku, nawet gdy wystąpi stan rzeczy zgodny z tym, który reprezentowany jest w zamiarze, to i tak zamiar nie zostanie uznany za spełniony. Niedopasowanie czasowe zamiaru do zachowania powoduje, że dane działanie nie jest traktowane przez sprawcę jako intencjonalne. Tylko zamiary są stanami wymagającymi tego typu synchronizacji czasowej. Tego typu dopasowanie nie jest wymagane w przypadku innych stanów intencjonalnych towarzyszących działaniom. Przykładowo, pragnienie zostanie zrealizowane, gdy zostaną spełnione warunki związane z jego treścią bez względu na to, czy stanie się to w momencie X, czy Y. Jeśli pragnąłem zobaczyć na żywo wodospad Niagara, to pragnienie to zostanie spełnione, gdy znajdę się w miejscu, z którego mógłbym podziwiać wspomniany wodospad. Natomiast nie będzie miało znaczenia to, co spowodowało, że mogłem go zobaczyć: czy jest to efekt moich starannie zaplanowanych działań, czy też znalazłem się w jego pobliżu przypadkiem.

Jaką zatem postać przybiera proces realizacji pełnego działania intencjonalnego, które jest zgodne z zamiarem? Zadaniem Searle'a powinno ono przebiegać według następującego schematu:

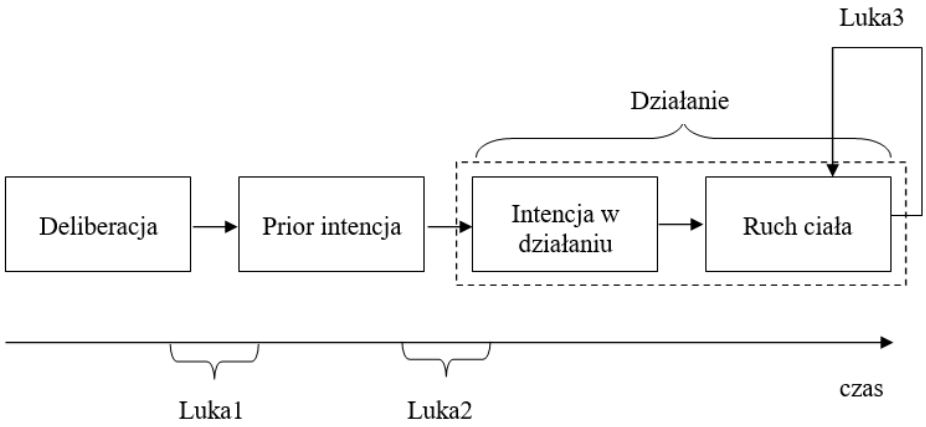


Diagram 3. Schemat przebiegu działania intencjonalnego wg Searle'a (Searle, 1983, s. 98).

Diagram 3 przedstawia kluczowe etapy realizacji pełnego działania intencjonalnego. Strzałki na diagramie reprezentują relację przyczynową między poszczególnymi komponentami działania. Równocześnie, nie należy traktować wskazanej sekwencji jako typowego ciągu przyczynowo-skutkowego, w którym wystąpienie jednego składnika prowadzi z koniecznością do pojawienia się drugiego – na tę szczególną cechę działań intencjonalnych wskazują luki z odpowiednimi indeksami (1, 2, 3).

Deliberacja

Pierwszy etap – deliberacja – to realizacja procesu decyzyjnego. W jego trakcie wybrane zostaje zachowanie (lub sekwencja zachowań) niezbędne do osiągnięcia obranego przez podmiot celu. Proces deliberacji jest, jak twierdzi Searle, specyficznego rodzaju działaniem, przy czym jego rezultatem nie jest zachowanie, a tzw. prior intencja, czyli szczególnego rodzaju stan intencjonalny. Podstawą deliberacji jest sieć stanów intencjonalnych. Wykonywane w jej ramach operacje umysłowe prowadzą do podjęcia decyzji zapamiętanej przez umysł w formie zamiaru (prior intencji). Jak tego typu operacje przebiegają? Jakie czynniki wpływają na podjęcie danej decyzji? Różne dyscypliny naukowe (m.in. ekonomia, psychologia czy filozofia) starają się odpowiedzieć na powyższe (i zbliżone do nich) pytania. Searle nie poświęca zbyt wiele uwagi deliberacji w

eseju na temat intencjonalności[20], traktując ją jako zjawisko oddzielne, wymagające odrębnych badań, niezależnych od rozważań nad naturą działania intencjonalnego. Podobne ograniczenie zakresu analiz przyjąłem w niniejszej pracy.

Prior intencja

Prior intencja to stan, który zgodnie z propozycjami terminologicznymi Searle'a, posiada tzw. niezerowe nakierowanie na zgodność typu: świat → umysł, to znaczy: składające się na niego warunki spełniania zostaną z sukcesem zrealizowane wówczas, gdy (1) stosowny fragment świata „zostanie dopasowany" do umysłu oraz (2) działanie przebiegnie zgodnie z przedstawionym powyżej schematem. Oczywiście, „dopasowanie świata do stanu umysłu" jest wyrażeniem o charakterze metaforycznym, gdyż w praktyce inicjatywa należy do podmiotu, który poprzez swoją aktywność dąży do wprowadzenia zmiany w świecie zgodnej z warunkami spełniania prior intencji. Od strony treściowej przedmiotem prior intencji (p) jest określonego rodzaju zachowanie, np. „pojadę dzisiaj do kina", „zadzwonię do przyjaciela", „napiszę kolejny fragment artykułu", „zapiszę się do klubu żeglarskiego". Warto zwrócić uwagę, że na tym etapie mówimy nie o jednostkowym działaniu (*token*), ale o pewnym typie (*type*) działania, które „zostanie" skonkretyzowane w momencie jego realizacji.

Prior intencji odpowiada od strony językowej następujące wyrażenie: „zrobię/wykonam A" albo „Zamierzam zrobić/wykonać A" (*I will do A* albo *I'm going to do A*). Istotne jest również to, że treść prior intencji jest na ogół dość abstrakcyjna i wysokopoziomowa. Nie uwzględnia ona szeregu działań pomocniczych koniecznych do jej realizacji, takich jak np. ruchy ręką niezbędne do przełączania biegów w samochodzie czy otwarcie drzwi podczas wizyty w sklepie spożywczym. W takich przypadkach niemal cała uwaga skupiona jest nie na zachowaniach, ale na ich oczekiwanych rezultatach.

Jak już wspominano, każdy tego typu zamiar, zgodnie z zasadą holizmu znaczeniowego, osadzony jest w szerszym kontekście intencjonalnym. W myśl tej zasady, zamiar pójścia do kina nie mógłby powstać w umyśle średniowiecznego mnicha, gdyż warunkiem pojawienia się tego typu stanu jest odpowiednio rozbudowana sieć przekonań, która obejmuje zarówno środowisko społeczne, jak i fizyczne. Na przykład, podmiot, którego intencją jest pójście do kina powinien: (1) wiedzieć, że kino to pewna instytucja

[20] Deliberacja, czyli m.in. podejmowanie racjonalnych wyborów, jest przedmiotem badań omówionych w innej pracy Searle'a zatytułowanej *Rationality in action*, która została wydana w 2001 roku.

kulturalna funkcjonująca według ściśle określonych reguł, (2) umieć się do nich stosować, a przede wszystkim (3) posiadać kompetencje niezbędne do odbioru dzieła filmowego.

Główną funkcją prior intencji jest dostarczenie reprezentacji działań, które mają doprowadzić do osiągnięcia celu wybranego w toku deliberacji. W przypadku ludzkiego działania taki stan intencjonalny jak prior intencja może obejmować bardzo długi horyzont czasowy. Większość naszych planów dotyczy najbliższych minut, godzin, dni lub tygodni, ale mamy również zamiary długoterminowe, takie choćby jak dążenie do zdobycia wykształcenia. Łatwo dostrzec, że tego typu zamiary wiążą się w pewne hierarchie i wzajemnie się uzupełniają. Plany wyznaczone przez prior intencję zwykle nie są szczegółową specyfikacją poszczególnych zachowań. Zazwyczaj treść tych stanów zawiera jedynie listę najważniejszych działań oraz charakterystykę oczekiwanych efektów ich realizacji. Postawić można pytanie: jak tego typu reprezentacje pomagają w skutecznym osiąganiu celu? Searle przedstawił pewne wskazówki na ten temat, kiedy rozważał kwestię złożonych działań intencjonalnych. Prezentacją tego zagadnienia zajmę się w kolejnym podrozdziale.

Intencja w działaniu

Zaprezentowany wyżej Diagram 3. pokazuje, że podczas realizacji pełnego działania intencjonalnego prior intencja nie wpływa bezpośrednio na sam ruch ciała, ale oddziałuje przyczynowo na całość złożoną z dwóch komponentów (1) intencji w działaniu oraz (2) ruchu ciała (zachowania). Znaczy to, że realizacja planu (prior intencji) wymaga nie tylko odpowiedniego ruchu ciała, ale również powiązanego z nim specyficznego stanu intencjonalnego, który Searle nazywa intencją w działaniu albo poczuciem działania (*experience of acting*).

Intencja w działaniu, podobnie jak prior intencja, charakteryzuje się nakierowaniem na zgodność typu: świat→umysł, tzn., że ten intencjonalny stan umysłowy zostanie spełniony, o ile w świecie zajdą określone zmiany. Ponadto, mając na uwadze przyczynowy aspekt intencji, Searle wyróżnia w jej treści zbiór warunków spełniania odpowiedzialnych za realizację zachowania niezbędnego do przekształcenia zamiaru w oczekiwany stan świata. Zgodnie z teorią intencjonalności, tak określoną strukturę warunków spełniania cechuje tzw. samoodniesienie przyczynowe. Znaczy to, że warunki spełniania, które prezentowane są w treści stanu, nie tylko reprezentują przewidywane zmiany w świecie lub planowane ruchy ciała, ale także są odpowiedzialne za aktywowanie określonych programów

motorycznych w mózgu, które prowadzą do zaistnienia tychże ruchów i w konsekwencji powodują pojawienie się określonej zmiany w świecie. Innymi słowy, część ze wskazanych warunków spełniania reprezentuje to, co nastąpi, a część pełni funkcję przyczyny w odniesieniu do realizowanego działania. Od strony formalnej intencję w działaniu można wyrazić za pomocą złożenia następujących przekonań i pragnień:

„**Intencja** (zrobię A) →

Istnieje pewien stan intencjonalny x taki, że x zawiera:

- **Przekonanie** (możliwe jest, że zrobię A) oraz
- **Pragnienie** (zrobię A) oraz
- **Przekonanie** (x będzie działać przyczynowo, by wytworzyć: zrobię A) oraz
- **Pragnienie** (x spowoduje [przyczynowo – uwaga M.C.]: zrobię A)"[21] (Searle, 1983, s. 104).

Dwa pierwsze warunki odnoszą się do postaw wobec działania, natomiast dwa ostatnie mają charakter samoodniesień. Searle zwraca uwagę na fakt, że samoodniesienie powoduje, iż pomiędzy intencją w działaniu a ruchem ciała powstaje szczególnego rodzaju więź. Kiedy podnoszę rękę, gdy chcę zagłosować lub obracam się, by sprawdzić, kto mnie woła, złożona z treści i zachowania całość prezentuje mi się w sposób natychmiastowy i w pełni zsynchronizowany z przebiegiem zachowania. W ten sposób mogę odpowiedzieć bez zbędnej zwłoki na pytanie: „co teraz robisz?". Jest to, zdaniem Searle'a, sytuacja analogiczna do przeżycia percepcyjnego, które swoją treść intencjonalną prezentuje nam w sposób natychmiastowy i bezpośredni. Różnica pomiędzy prior intencją a intencją w działaniu polega na tym, że pierwsza odnosi się do działania w jego zgrubnej postaci, natomiast druga odnosi się do działania ukonkretnionego i realizowanego w danym momencie w formie określonego programu motorycznego. Zilustrować to można następującym przykładem. Mam zamiar pozdrowić znajomego, który stoi po drugiej stronie ulicy i patrzy w moim kierunku. Zrealizować to mogę na różne sposoby: mogę pomachać mu ręką, ukłonić się czy krzyknąć „Cześć" w jego kierunku. Mój uprzedni zamiar nie specyfikuje konkretnej formy intencjonalnego pozdrowienia, jednak bez takiego

[21] "Int (I will do A) →
There is some intentional state x such that x contains
Bel (◊I will do A) &
Des (I will do A) &
Bel (x will function causally toward production of: I will do A) &
Des (x will cause: I will do A)" (Searle, 1983, s. 104).

zamiaru nie wykonałbym żadnego z tych działań. Searle, rozważając tego typu kwestie, wyraźnie akcentuje doniosłość przyczynowych zależności pomiędzy prior intencją, intencją w działaniu oraz ruchem. W jego opinii, działanie nie może zostać uznane za zrealizowane zgodnie z planem, o ile to prior intencja, w szczególności część jej warunków spełniania, nie wywoła tego typu działania. Dany ruch nie może zostać uznany za zamierzony, o ile nie zostanie on wywołany przez intencję w działaniu. W tym przypadku, decydujące są warunki spełniania, cechujące się samo-odniesieniem przyczynowym. Znaczy to, że w złożonej sytuacji mamy do czynienia z następującą relacją: prior intencja S1(p1) przyczynowo wywołuje intencję w działaniu S2(p2), a ta z kolei prowadzi przyczynowo do zachowania Z. Co więcej, Searle zakłada przechodniość relacji przyczynowości. Wskazany ciąg przyczynowy rozpoczyna się na poziomie stanów intencjonalnych, a kończy na zachowaniu i wywołanych przez nie zmianach w świecie.

Dane empiryczne potwierdzają, w opinii Searle'a, istniejący w działaniu intencjonalnym ścisły związek między komponentem intencjonalnym a behawioralnym. Aby to zilustrować, omawia on dwa eksperymenty. Pierwszy – opracowany przez Williama Jamesa – ma następujący przebieg: badany proszony jest o zamknięcie oczu i podniesienie ręki. Wcześniej jednak ręka, która ma zostać podniesiona, zostaje znieczulona. Gdy uczestnik zamknie oczy, eksperymentator przytrzymuje rękę i prosi uczestnika o jej podniesienie. W momencie, gdy uczestnik otwiera oczy, pojawia się zaskoczenie, gdyż – wbrew oczekiwaniom – ręka nie znajduje się nad głową, lecz zwisa wzdłuż tułowia (James, 1950, s. 489). Efekt zdziwienia potwierdza, zdaniem Searle'a, że przeżycie fenomenalne towarzyszące intencji w działaniu może zostać odseparowane od zachowania.

Odwrotny efekt przedstawiony jest w eksperymencie Penfielda. Podczas operacji na otwartym mózgu – z zastosowaniem znieczulenia miejscowego oraz przy pełnej świadomości pacjentów – neurochirurg wywoływał u nich, poprzez odpowiednią stymulację kory ruchowej, niewielkie ruchy ręką lub krótką wokalizację (Penfield, 1975, s. 76). Następnie pacjenci pytani byli o ich stosunek do wywołanego ruchu lub wokalizacji. Wszyscy zgodnie raportowali, że nie mieli nic wspólnego z wywołanym zdarzeniem, a także prawidłowo rozpoznali, że przyczyną zachowania są manipulacje Penfielda. Eksperyment ten dobrze pokazuje, jak istotny jest komponent intencjonalny. Bez niego zachowania, które zewnętrznemu obserwatorowi jawią się jako intencjonalne, nie są takimi dla osoby je wykonującej. Dla niej są one co najwyżej odruchami, tikami albo ruchami wywołanymi przez czynnik zewnętrzny, np. wtedy, kiedy ruch jej ręki jest skutkiem

potrącenia przez osobę trzecią. W takich przypadkach określenia stosowane w odniesieniu do działań intencjonalnych, takie jak: „próbowałem ...", „udało mi się ...", „nie udało mi się ..." nie mogą być użyte, gdyż odnoszą się one do procesów umysłowych (prior intencja, intencja w działaniu) poprzedzających zachowanie i traktowanych przez podmiot jako jego przyczyna. Kiedy ktoś obserwuje i odczuwa ruchy własnego ciała, lecz nie ma żadnych „wskazówek" od swojego umysłu, że są one przez niego zamierzone i zaplanowane, nie może mówić o nich z perspektywy pierwszoosobowej, że to on wprawił swoje ciało w ruch.

Ktoś mógłby zapytać: dlaczego wyjaśnianie działań intencjonalnych wyłącznie za pomocą prior intencji oraz zachowania jest niewystarczające? Searle podaje kilka powodów. Pierwszy z nich odnosi się do rozróżnienia działań zaplanowanych i spontanicznych. Działania zaplanowane wymagają zrealizowania pełnego schematu, zaprezentowanego powyżej (Diagram 3.). W pierwszym kroku realizowany jest proces deliberacji, który kończy się ustanowieniem prior intencji. Następnie, kiedy pojawią się sprzyjające okoliczności, prior intencja powoduje, że podjęte zostanie określone działanie. Najpierw, zgodnie z mechanizmem przedstawionym na powyższym diagramie, pojawi się intencja w działaniu, która wywoła określoną sekwencję zachowań. Działania spontaniczne, w przeciwieństwie do zaplanowanych, nie posiadają fazy deliberacji, a ich zamierzony charakter nadaje im wyłącznie intencja w działaniu, która odpowiada za „uruchomienie" odpowiedniej do danej sytuacji sekwencji zachowań. Na przykład, dzieje się tak wtedy, gdy stojąc w towarzystwie komentujemy cudze wypowiedzi lub żartujemy. W takich przypadkach o treści naszych wypowiedzi decydują „luźne" skojarzenia, przypomniane historie, wcześniejsze doświadczenia, itp. Tego typu konwersacje rzadko bywają planowane, nawet jeśli są wysoce skonwencjonalizowane. Równocześnie, ich spontaniczność jest wyraźnie ograniczona przez bieżący kontekst np. ostatnie wydarzenia polityczne, kulturalne, plotki itd. Innym przykładem działań spontanicznych mogą być „pojedynki" zespołów kabaretowych. Na początku takiego wydarzenia uczestnikom prezentowany jest tzw. temat przewodni oraz obsadzane są poszczególne role. Na tej podstawie, bez żadnego przygotowania komicy/standup'erzy mają wykreować szereg zabawnych dialogów i scenek. Z perspektywy opracowanego przez Searle'a schematu mamy tu do czynienia z działaniami, którym brakuje fazy deliberacji, a co za tym idzie – również prior intencji. Nadal jednak każdy z uczestników wydarzenia będzie miał poczucie, że pomimo wielu

nieprawdopodobnych zwrotów akcji, wszystkie wypowiedziane kwestie były w pełni intencjonalne. Za tego typu efekt, zdaniem Searle'a, odpowiada intencja w działaniu.

Drugi argument odwołuje się do analizy przypadków szczególnych, przedstawionych przez filozofów umysłu i działania[22] w formie eksperymentów myślowych. Searle omówił między innymi eksperyment myślowy Davidsona traktujący o dwóch alpinistach, z których jeden zwisa na linie asekuracyjnej, a drugi podtrzymuje go, by nie doszło do wypadku. W umyśle alpinisty trzymającego linę rodzi się chęć, pod wpływem ciężaru i w obliczu narastającego niebezpieczeństwa, by puścić tę linę i w ten sposób ocalić przynajmniej siebie. Kiedy alpinista zaczyna pojmować, z jakimi konsekwencjami dla partnera wiąże się jego ewentualna decyzja, to jego myśli opanowuje mimowolna nerwowość, która prowadzi do rozluźnienia uchwytu i w konsekwencji do wypuszczenia liny[23] (D. Davidson, 1973, s. 153). Searle zauważa, analizując opracowany przez Davidsona przykład, że gdyby nawet uznać pragnienie, które zrodziło się w umyśle alpinisty, za prior intencję, to już jej realizację trudno potraktować jako zamierzoną. Innymi słowy, pragnienie zrealizowania działania nie jest warunkiem wystarczającym, by zachowanie zgodne z pragnieniem mogło być utożsamione przez jego sprawcę z działaniem intencjonalnym. Aby puszczenie liny miało status działania intencjonalnego, alpinista musiałby odpowiedzieć na pytanie: „Co teraz robisz?" (*what are you now doing?* (Searle, 1983, s. 90)) – „Staram się uwolnić od ciężaru, puszczając linę". Natomiast w sytuacji, w której przyczyną jego zachowania jest mimowolna nerwowość, odpowiedziałby: „Staram się przytrzymać linę, która wymyka mi się z rąk.". Widać zatem, że intencja w działaniu, prezentująca i uruchamiająca zachowanie (ruch), jest istotnym i niezbędnym składnikiem pełnego działania intencjonalnego.

Kontynuacja działania

Searle w zaproponowanym schemacie omówił również mechanizm odpowiedzialny za kontynuację działania intencjonalnego. Niestety, składnik ten potraktowany został przez filozofa zdawkowo i nie wiadomo, jak należałoby go rozumieć. Nie jest na przykład jasne, czy tego typu kontynuacja dotyczy wyłącznie aktualnie realizowanej intencji w działaniu,

[22] Ciekawe eksperymenty myślowe tego typu zaproponowali: R.M. Chisholm (Chrisholm, 1966), D. Davidson (D. Davidson, 2001), D. Bennett (w: Davidson, 2001).

[23] Davidson przedstawia ten eksperyment następująco: „Posłużmy się tylko jednym przykładem. Wspinacz mógłby chcieć pozbyć się ciężaru i niebezpieczeństwa związanego z trzymaniem liny z wiszącym na niej kolegą. **Przekonanie** to i ta **chęć** mogłyby go tak zdenerwować, że rozluźni on chwyt na linie, pomimo, że tak naprawdę nigdy nie zdecydował się, by rozluźnić chwyt ani nie zrobił tego intencjonalnie (tłum. własne)". (D. Davidson, 2001, s. 153).

czy poprzedzającego ją planu. Wydaje się, że Searle traktował tę kwestię jako drugorzędną i dlatego nie poświęcił jej należytej uwagi.

Z perspektywy filozoficznej, podstawowej dla Searle'a, zagadnienie kontynuacji działań intencjonalnych jest być może mniej istotne, jednak gdy celem analizy jest konstrukcja modelu obejmującego najważniejsze mechanizmy organizujące złożone sekwencje zachowań, to ma ono niebagatelne znaczenie. W ostatnim rozdziale rozprawy zaproponuję charakterystykę mechanizmu konstruującego sekwencje zachowań oraz wskażę, jak można włączyć go w model wyjaśniający przebieg złożonego działania intencjonalnego.

Luka

Ostatnim elementem, który dopełnia charakterystykę działań intencjonalnych w schemacie Searle'a, jest tzw. luka (*gap*). Luka to puste miejsce, czy też odstęp pomiędzy składnikami umysłowymi działania intencjonalnego[24]. To luki sprawiają, że agent zachowuje poczucie, iż to on – a nie okoliczności względem niego zewnętrzne – decyduje o tym, że jego działanie jest zgodne z jego własnym, swobodnie przez niego wybranym, zamiarem. Searle wyjaśnia, że ma na myśli sytuacje, w których determinanty powiązane z działaniem **nie są** doświadczane przez sprawcę jako warunki wystarczające do jego realizacji (Searle, 2001, s. 269). W rezultacie odnosimy wrażenie, że „mogliśmy postąpić inaczej", niż faktycznie postąpiliśmy (*I could have done otherwise*) (Harris, 2012, s. 39). Wybraliśmy A, ale równie dobrze w tych samych okolicznościach mogliśmy wybrać B. Zrealizowaliśmy działanie C, choć mogliśmy także wykonać działanie D. Wskazane alternatywy odnoszą się nie tylko do przypadków brzegowych (gdy wahamy się pomiędzy dostępnymi możliwościami), ale także do sytuacji, w których nasz wybór jest zdecydowany i najprawdopodobniej w podobnych okolicznościach zostałby powtórzony. Nawet wtedy mamy poczucie, że to od nas zależało, które działanie podejmiemy. Choć podjęliśmy taką, a nie inną decyzję, to mogliśmy dokonać innego wyboru i zaplanować wszystko inaczej. W przebiegu działania intencjonalnego wyróżnić można, zdaniem Searle'a, trzy tego typu luki. Pierwsza pojawia się po prior intencji, druga po intencji w działaniu, a trzecia odnosi się do mechanizmu odpowiedzialnego za kontynuowanie działania. Searle utrzymuje, argumentując na rzecz trafności ujęcia działania intencjonalnego jako procesu zawierającego luki, że ich

[24] "«The gap» is the general name that I have introduced for the phenomenon that we do not normally experience the stages of our deliberations and voluntary actions as having causally sufficient conditions or as setting causally sufficient conditions for the next stage." (Searle, 2001, s. 50).

występowanie jest wyzwaniem dla typowych analiz naukowych i filozoficznych odwołujących się do związków przyczynowych. Zgodnie ze standardowym pojmowaniem takich związków, jeśli jedno zjawisko zostanie zidentyfikowane jako przyczyna innego, to nie sposób sobie wyobrazić, by to pierwsze wywołało inny skutek, niż ten, który faktycznie się pojawił. Gdy widzimy efekt w postaci trzęsienia ziemi i wiemy, że na obszarze tym zarejestrowano ruchy płyt tektonicznych, to nie sposób przyjąć, że ich przesunięcie nie wywołało trzęsienia. Tymczasem, w przypadku działania intencjonalnego jeden stan umysłowy może prowadzić do różnych działań. Można zatem powiedzieć, że przyczynowość, która związana jest z działaniami intencjonalnymi, pozostaje „nieciągła", czyli posiada luki. W filozofii tę szczególnego rodzaju „swobodę" i związany z nią brak konieczności kauzalnej zwykło się określać mianem problemu wolnej woli. Searle twierdzi, że dotychczas problem ten nie uzyskał satysfakcjonującego rozwiązania, pomimo wielowiekowej refleksji i wielkiej różnorodności stanowisk (Searle, 2010a).

2.3 Proste i złożone działania intencjonalne

Przedstawiona wyżej koncepcja Searle'a jest na tyle ogólna, że odnosi się zarówno do prostych, jak i złożonych działań intencjonalnych. Jej autor zaznacza równocześnie, że zaproponowany przez niego schemat, w przypadku działań złożonych, wymaga uzupełnienia o dodatkowe elementy. Według Searle'a, w przypadkach intencjonalnych działań prostych, treść intencji odnosi się głównie do projektowanego zachowania oraz oczekiwanej zmiany w świecie i poza te reprezentacje w zasadzie nie wykracza. Na ogół jednak nasze intencje mają postać złożoną, np. „zamierzam za 2 dni polecieć na urlop do Pekinu" lub „zamierzam dać im do zrozumienia, że nie należy tak postępować". W tego typu przypadkach czysto „zachowaniowy", czyli motoryczny wymiar działania jest tylko jednym z elementów całego zbioru warunków spełniania wchodzących w skład prior intencji. Tak pojmowana prior intencja jest *de facto* kompleksem złożonym z wielu składników. Mogą to być intencje składowe, niezbędne do zrealizowania intencji naczelnej, mogą to także być stany umysłowe odnoszące się do uwarunkowań historyczno-geograficzno-cywilizacyjnych, ale także stany fizyczne organizmu (w pierwszej kolejności – mózgu) podbudowujące jego zdolność do generowania stanów umysłowych, a nawet stany otoczenia. Zilustrujmy to na przykładzie przytoczonego wyżej przykładu. Kiedy żywię zamiar, aby za 2 dni polecieć na urlop do Pekinu, to nie jest to czyste pragnienie,

aby za 2 dni znaleźć się w Pekinie[25], ani prosta prior intencja, bo ta odsyłałaby jedynie do zachowań motorycznych. Jego zrealizowanie wymaga stworzenia realistycznego – choć bardzo prowizorycznego – planu postępowania, a także wystąpienia szeregu niezbędnych stanów świata. Pojawienie się tych stanów zależy od pojawienia się stosownych warunków: historycznych (muszą to być czasy, w których odbywają się powietrzne, pasażerskie loty do Pekinu), geograficznych (w miejscu, z którego rozpocznę podróż, jest odpowiednie lotnisko, z którego można odprawić się do Pekinu) i cywilizacyjnych (ich lista jest bardzo obszerna, ale łatwo wskazać te, które taką podróż umożliwiają, np. tablice informacyjne, automaty służące do rezerwacji miejsc w samolocie, stanowiska odpraw, aplikacje mobilne przechowujące bilety lotnicze, itp.). Natomiast plan postępowania zawiera szkic listy działań, zarówno umysłowych, jak i czysto fizycznych, które należy wykonać, aby doprowadzić do zrealizowania zamiaru. Działania te również są uwarunkowane przez stany ciała oraz stany otoczenia. Są wśród nich stany, których agent nie wziął albo wręcz nie mógł wziąć pod uwagę. Czasami są to dodatkowe aspekty działania, np. wpływ poszczególnych procesów mózgowych na przebieg działania, niekiedy zaś są to efekty niezamierzone, tzw. czynniki uboczne. Łatwo zauważyć, że tego rodzaju poszerzenie charakterystyki działania intencjonalnego bardzo je rozmywa oraz sprawia, że jego opis w znacznym stopniu traci przydatność do objaśniania ludzkiej aktywności. Widać to choćby w przypadku zaproponowanego przez Searle'a rozszerzania, które nazywa on **efektem akordeonu** (Searle, 1983, s. 98).

Przykładem, którym posługuje się Searle do zobrazowania struktury złożonej intencji, jest tragiczne zdarzenie z historii, a mianowicie zabójstwo arcyksięcia Franciszka Ferdynanda. Z perspektywy zabójcy, serbskiego nacjonalisty Gavrilo Principa, to wydarzenie można by sprowadzić do następujących etapów:

1. pociągnął za spust,
2. wystrzelił z pistoletu,
3. postrzelił arcyksięcia,
4. zabił arcyksięcia,
5. zadał cios Austrii,
6. pomścił Serbię.

[25] Przypomnę, że do warunków spełnienia pragnienia nie należy aktywność jego podmiotu, ale wystarczy, aby zrealizował się stan świata, będący odniesieniem przedmiotowym tego stanu intencjonalnego.

Pozycje 1-4 uporządkowane są zgodnie z relacją przyczynową, np. wystrzelenie z pistoletu zrealizowane zostało „za pomocą" (*by means of*) pociągnięcia za spust. Z kolei między pozycją 5 i 4 oraz 6 i 4 zachodzi relacja konstytuowania. Oznacza to, że zadanie ciosu Austrii lub pomszczenie Serbii zrealizowane zostało „przez" (*by way of*) zabicie arcyksięcia. Dodatkowo, powyższą listę można uzupełnić o następujące pozycje, mając na uwadze efekt akordeonu:

1. aktywował neurony w korze motorycznej odpowiedzialne za skurcze mięśni w jego ramieniu oraz dłoni,
2. pociągnął za spust,
3. wystrzelił z pistoletu,
4. spowodował ruch wielu cząstek w powietrzu,
5. postrzelił arcyksięcia,
6. zabił arcyksięcia,
7. zadał cios Austrii,
8. pomścił Serbię,
9. zepsuł lato lordowi Gray'owi,
10. przekonał Cesarza Franciszka Józefa, że Bóg ukarał rodzinę cesarską,
11. rozwścieczył Wilhelma II,
12. rozpoczął Pierwszą Wojnę Światową.

Punkty 1 i 4 uzupełniają sekwencję przyczynową o dodatkowe ogniwa, ale same w sobie nie muszą być częścią intencji. Podmiot nie musi być ich świadomy, by doszło do realizacji zamiaru. Z kolei pozycje 9-12 są przykładami rezultatów ubocznych powiązanych z działaniem, ale bez związku z intencją (są to tzw. nieintencjonalne aspekty działania). Widać zatem, że efekt akordeonu, tj. możliwość włączania w opis działania intencjonalnego elementów uszczegółowiających lub rozszerzających, może prowadzić do nieuprawnionego przypisania sprawcy niezamierzonych przez niego efektów działania. O ile z pewnym prawdopodobieństwem można przyjąć, że Gavrilo Princip chciał pomścić Serbię, o tyle trudno uznać, że planował przez swój terrorystyczny akt zepsuć lato lordowi Gray'owi. Jeśli jednak na efekt akordeonu nałoży się treść złożonej intencji, to – zdaniem Searle'a – nie musi to prowadzić do nieuprawnionych nadinterpretacji. Istotne okazuje się to, że przywołany efekt jest przejawem szczególnej zdolności gatunku ludzkiego do stopniowego przesuwania uwagi z aspektu motorycznego zachowań na ich aspekty

funkcjonalne, w tym także znaczenia, konstytuowane za pośrednictwem działań w sferze społecznej.

> *Princip poruszył tylko palcem, ale jego intencjonalność objęła całe Imperium Austro-Węgierskie. Tego typu zdolność do dysponowania warunkami spełniania wykraczającymi poza ruchy ciała to klucz do zrozumienia znaczenia oraz przyczynowości.*[26]

Przywołany wyżej aparat pojęciowy w przejrzysty i intuicyjny sposób określa – w języku Searle'owskiej teorii intencjonalności – najważniejsze składowe prostego działania intencjonalnego (deliberacja, prior intencja, intencja w działaniu, zachowanie, mechanizm kontynuacji, luki) oraz istniejące między nimi relacje (związek przyczynowy między prior intencją a działaniem oraz związek przyczynowy pomiędzy intencją w działaniu a zachowaniem). Skoro omówiony schemat (patrz: Diagram 3.**Error! Reference source not found.**) odnosi się tylko do prostego działania, to pojawia się pytanie: jak należy zmodyfikować tenże schemat, aby można go było odnieść do działań złożonych? Zdaniem Searle'a, należy przede wszystkim odsłonić i opisać wewnętrzną strukturę złożonej prior intencji. Pierwszym krokiem w tym kierunku jest charakterystyka efektu akordeonu i wskazanie, jak jego uwzględnienie wpływa na modyfikację modelu prostego działania intencjonalnego. Następnym krokiem jest uwzględnienie sposobu hierarchizowania poziomów intencji i – co za tym idzie – określenia natury działań podstawowych (Searle, 1983, s. 98). Treść poniższego podrozdziału odnosi się do tego zagadnienia.

2.4 Działania podstawowe

Z efektem akordeonu wiąże się jeszcze jeden intrygujący problem, mianowicie kwestia tzw. działań podstawowych. Zgodnie z definicją Searle'a:

> *A jest działaniem podstawowym dla agenta S wtedy i tylko wtedy, gdy (1) S jest w stanie zrealizować akt typu A oraz (2) S może posiadać zamiar zrealizowania aktu typu A bez odwoływania się do innych działań, by móc zrealizować akt A.* (Searle, 1983, s. 100).

[26] "Princip moved only his finger but his Intentionality covered the Austro-Hungarian Empire. This capacity to have additional conditions of satisfaction beyond our bodily movements is a key to understanding meaning and causation" (Searle, 1983, s. 99).

Innymi słowy, działanie podstawowe to takie, które jest przez sprawcę postrzegane jako tak proste, że – aby je wykonać – nie musi rozkładać go na składowe działania intencjonalne, które musiałby realizować po kolei.

Działaniem podstawowym dla Gavrilo Principa było pociągnięcie za spust (czyli pierwsza pozycja na liście warunków spełniania). Gdyby Gavrilo był snajperem, to prawdopodobnie jego intencja miałaby postać: „strzelić w klatkę piersiową". Wskazana różnica między obiema intencjami wiąże się ze specyficzną cechą działań podstawowych, jaką jest ich relatywizacja do podmiotu. Innymi słowy, dla każdego podmiotu wyróżnić można specyficzną dla niego klasę działań podstawowych zależną od jego indywidualnych doświadczeń i wyuczonych zachowań. Zdolność do organizowania sekwencji zachowań w działania podstawowe to, zdaniem Searle'a, przejaw ogólnej tendencji umysłu do łączenia w jednostki wyższego rzędu tego, co powtarzalne i rutynowe. Tego typu jednostki składają się na ogół ze stanu intencjonalnego oraz umiejętności należących do tła. Działania podstawowe mogą przyjąć – po odpowiednim treningu – bardzo złożone formy. Przykładem może tu być doskonalenie umiejętności graczy komputerowych, dla których złożone sekwencje ruchów ręki i uderzeń palcami dłoni w przyciski sprowadzają się do pojedynczych działań intencjonalnych, takich jak: przeskoczyć przeszkodę wykorzystując energię eksplozji z granatnika, wykonać salto w powietrzu, itp. Zauważmy, że gracz początkujący skupiać będzie intencjonalną uwagę na doskonaleniu ruchów ręki, natomiast dla gracza zaawansowanego najważniejsze będzie to, co dzieje się na ekranie, a poczucie kontroli ruchu przenosi się z ręki i palców na kontrolę klawiszy czy też manipulatora do gry (Dayan & Cohen, 2011). Przekształcone do postaci działań podstawowych sekwencje zachowań umożliwiają skuteczniejsze osiąganie celów w niezwykle dynamicznym środowisku gry. Searle twierdzi, że systematyczny trening lub rutyna prowadzą do wykształcania się zupełnie nowej, „zoptymalizowanej" reprezentacji wielu zachowań składających się na dane działanie lub zbiór działań. Nie jest to zatem proces stopniowego „przesuwania" z pola świadomości do nieświadomości poszczególnych stanów intencjonalnych kontrolujących przebieg działań, ale kompleksowa przebudowa całej sekwencji zachowań.

Stopniowe przekształcanie wielokrotnie powtarzanych działań w jednostki wyższego rzędu stanowi jedną z ważniejszych cech ludzkiego systemu kontroli zachowań. Nie ulega wątpliwości, że w tworzenie takich jednostek zaangażowane są procesy uczenia się i optymalizowania działań, pomocne w coraz efektywniejszym wykorzystywaniu

posiadanych zasobów. Searle zdaje sobie z tego sprawę, jednak nie objaśnia – czy i jak należałoby włączyć procesy uczenia się w strukturę złożonego działania intencjonalnego. Równie ważną funkcją wskazanego mechanizmu jest hierarchizowanie zachowań, mające niebagatelne znaczenie dla procesów planowania. Złożone środowisko, w którym funkcjonuje człowiek, a takim jest jego otoczenie przyrodniczo-cywilizacyjne, wymaga odpowiednio zaawansowanego planowania. Gdyby nasze plany konstruowane były jedynie dla (całych sekwencji) prostych działań, to nie dość, że ich opracowanie byłoby bardzo żmudne i kosztowne, to równocześnie ich zakres nie przekraczałby prawdopodobnie kilkunastu czynności.

Zauważmy, że w koncepcji Searle'a działanie podstawowe może być zarówno proste, jak i złożone. Jednakże zostało ono wprowadzone przede wszystkim po to, aby pokazać, jak biegłe opanowanie pewnego kompleksu prostych działań sprawia, że nie są już one dla danego agenta sekwencją ruchów, z których do niedawna każdy był osobno kontrolowany, a stają się jednostkową czynnością, która nabiera nowego znaczenia. Wykonanie takiego działania wyższego rzędu wymaga realizacji działań składowych. Agent przenosi swoją uwagę na odpowiedni poziom i dokonuje redeskrypcji działania wymagającej zmiany zarówno prior intencji, jak i intencji w działaniu.

W ostatnim rozdziale pracy przedstawię model podsystemu planowania, który z jednej strony operuje jednostkami wyższego rzędu, a z drugiej – zarządza procesem stopniowej automatyzacji prostych działań.

2.5 Przyczynowy status intencji

Przedstawiona powyżej analiza, dotycząca poszczególnych typów działań, pokazuje, że związane z nimi typy intencji bezpośrednio oddziałują na przebieg zachowań. Warunki spełniania dla intencji – w odróżnieniu od tych dla pragnień, przekonań, wyobrażeń i innych czynności umysłowych – nie tylko reprezentują określone treści, ale również uczestniczą w procesie realizacji zamiaru. Jeśli sposób realizacji przebiegnie niezgodnie z treścią intencji, wówczas uzyskany efekt nie zostanie uznany za zgodny z zamiarem. Podmiot zachowania uzna je np. za przypadkowe, niezamierzone albo wymuszone przez czynnik zewnętrzny. W tego typu podejściu zakłada się, że intencja jest jednym z członów związku przyczynowego, którego skutkiem jest zachowanie. Jak zasygnalizowałem we

wprowadzeniu, takie ujęcie jest przedmiotem sporów w filozofii umysłu, który Julia Yoo relacjonuje w następujący sposób:

> *Filozofowie ciągle próbują nadać sens przyczynowości umysłowej. Wielu krytykuje założenia, na których rzekome problemy z przyczynowością umysłową są ufundowane, a w szczególności, zaproponowane przez Kima ujęcie problemu wykluczenia (Bennett 2003, Menzies 2003, Raymont 2003). Inni każą nam zaakceptować stanowiska, które już wcześniej zostały odrzucone jako nieprzydatne, jak np. fizykalizm typów (Hill, 1991), albo całkowicie nieużyteczne, jak np. epifenomenalizm (Bieri 1992, Chalmers 1996, rozdział 5).*
>
> *Niektórzy nawet kwestionują to, czy rzeczywiście mamy problem z przyczynowością umysłową (Baker 1993, Burge 1993). Baker (1993) argumentuje, że uznanie zasad fizykalizmu nie tylko obarcza nas problemem wykluczania, ale czyni go całkowicie nierozwiązywalnym. Jednocześnie, kontynuuje Baker, powszechny epifenomenalizm, który zapanowałby, gdybyśmy poważnie potraktowali zasady fizykalizmu, byłby równoznaczny z reductio ad absurdum samych tych zasad, dlatego musimy odrzucić te zasady, i w takim przypadku problem wykluczenia sam się rozwiąże. Baker dość radykalnie proponuje, by odrzucić zasadę domknięcia przyczynowego, o ile chcemy zachować możliwość przyczynowości mentalnej – w szczególności, jeśli chcemy zachować możliwość makro-przyczynowości – możliwość, która, jak twierdzi Baker, została z powodzeniem przetestowana w naszych praktykach wyjaśniających.*
>
> *Jednak Antony (1991) oraz Kim (1993) twierdzą, że problem przyczynowości mentalnej to w istocie problem wyjaśnienia, skąd się bierze sukces eksplanacyjny w przypadku wyjaśniania zachowań w języku wyrażeń mentalnych. Znaczy to, że problem nie zniknie przez stwierdzenie, że nasze mentalistyczne wyjaśnienia działają całkiem dobrze. Zagadką jest to, skąd bierze się skuteczność wyjaśnień mentalnych, skoro wszystkie argumenty metafizyczne wskazują na przyczynową nieistotność tego, co mentalne.*[27]

[27] "Philosophers are still busy at work trying to make sense of mental causation. Many criticize the assumptions on which the alleged problems of mental causation are predicated, particularly Kim's formulation of the exclusion problem (Bennett 2003, Menzies 2003, Raymont 2003). Others enjoin us to accept those very positions that have been cast aside as unavailable, such as type physicalism (Hill 1991), or down-right implausible, such as epiphenomenalism (Bieri 1992, Chalmers 1996, ch.5).

Some have even questioned whether we really have a problem concerning mental causation (Baker 1993, Burge 1993). Baker 1993 has argued that once the principles of physicalism are accepted, not only are we

W opinii Searle'a, przytoczone powyżej przez Yoo (2007) stanowiska opierają się na błędnym postrzeganiu relacji: stan mentalny – zachowanie. By lepiej zrozumieć stanowisko tego naturalisty biologicznego, warto przytoczyć jego najważniejsze argumenty dotyczące rozważanego zagadnienia.

Regularnościowa teoria przyczynowości

Kiedy podejmujemy różnego rodzaju działania, zazwyczaj przyjmujemy, że ich przyczyną jest poprzedzający je zamiar. Zdaniem Searle'a, problem polega na tym, że trudno jest „pogodzić" przyczynowy status intencji z dominującą w filozofii i nauce tzw. regularnościową teorią przyczynowości. Zgodnie z tą teorią związek przyczynowy jest relacją między dwoma zdarzeniami, która posiada następujące własności: (1) jest nieobserwowalna – obserwowalna jest tylko regularność, (2) podpada ona pod jakieś uniwersalne prawo przyrody, którego opis zawiera typy zdarzeń wchodzących w skład relacji przyczynowej, (3) odnosi się do zdarzeń niezależnych logicznie, a zatem do prawd przygodnych (niekoniecznych w sensie logicznym). Wszystkie wymienione cechy, zdaniem Searle'a, stają się problematyczne, gdy uwzględnia się przyczynowość intencjonalną.

Przede wszystkim, zdaniem amerykańskiego filozofa, teoria regularności przeczy naszym codziennym doświadczeniom, gdyż podważa, bezpośrednio dane nam poczucie, dostępne bez dodatkowych obserwacji, że realizowane przez nas działania zmieniają otaczające środowisko zgodnie z naszymi zamiarami lub dążeniami. Kiedy czuję pragnienie i w związku z tym napiję się wody, dokładnie wiem, dlaczego tak postąpiłem. Wiem, podnosząc rękę, jaka intencja stała za tym zachowaniem. Nie muszę analizować wcześniejszych przypadków ani dysponować wiedzą o ogólnych prawach przyrody, by wyjaśnić własne zachowanie. Wiem również, bez znajomości jakiejś uniwersalnej korelacji, że prawdziwe jest twierdzenie kontrfaktyczne głoszące, że gdybym nie był

saddled with the exclusion problem, but the problem is also absolutely unsolvable. But, Baker continues, the wide-scale epiphenomenalism that would ensue were we to take the principles of physicalism seriously is tantamount to a reductio ad absurdum of the principles themselves, so we must reject the principles, in which case the exclusion problem dissolves of itself. Baker quite radically proposes that we reject the causal closure thesis if we wish to hold onto the possibility of mental causation – indeed, if we want to hold onto the possibility of macro-causation generally – a possibility that Baker claims is well testified by the successes of our explanatory practices.

Antony 1991 as well as Kim 1993, however, have argued that the problem of mental causation is the problem of explaining how and why there is this explanatory success when it comes to explaining behavior in mental terms. That is, the problem does not go away by pointing out that our mentalistic explanations perform quite well. The puzzle is how they explain so well, given that the metaphysics all point to the causal irrelevance of the mental" (Yoo, 2021).

spragniony, to nie napiłbym się wody. Na niższym poziomie organizacji – w tego typu przypadkach – działają również niskopoziomowe prawa fizyczne wyjaśniające procesy istniejące w ciele oraz mózgu, jednak ich znajomość nie jest niezbędna, by wiarygodnie wyjaśnić przyczynę działania. Searle twierdzi, że nawet gdyby człowiek znał niskopoziomowe, uniwersalne prawo wyjaśniające jego zachowanie, to nie miałoby ono dla niego takiego znaczenia, jak wyjaśnienie odwołujące się do bezpośredniej wiedzy o pragnieniu. Co więcej, w odróżnieniu od związków przyczynowych mających swoją podstawę w prawach przyrody, związki oparte na przyczynowości intencjonalnej nie są przez ludzi uznawane jako konieczne. Nawet, jeśli w podobnych okolicznościach człowiek ponownie napiłby się wody, to nadal miałby poczucie, że czyn ten zależy od jego osobistej, aktualnie podjętej decyzji, a nie od jakichś niezależnych od niego praw przyrody (Searle, 1983, s. 118). Widać zatem, że tego typu związki między zdarzeniami trudno ujmować jako szczególne przypadki uniwersalnych zależności. Dlatego też – przynajmniej w przypadku działań intencjonalnych – należy porzucić regularnościową teorię związku przyczynowego, gdyż nie uwzględnia ona doświadczanej przez człowieka przyczynowości intencjonalnej.

Bezpośredni charakter przyczynowości intencjonalnej

Skoro przyczynowość intencjonalna wymyka się opisowi regularnościowemu, to należy, zdaniem Searle'a, raz jeszcze przyjrzeć się pojęciu związku przyczynowego. Amerykański filozof proponuje uznać, w duchu naiwnego realizmu, że związek przyczynowy jest szczególnego rodzaju relacją zachodzącą między określonymi obiektami w świecie. Takie ujęcie, zdaniem Searle'a, znosi wymóg, by dany związek przyczynowy był szczególnym przypadkiem ogólnej prawidłowości mającej postać prawa przyrody. Pozwala ono także wyjaśnić następujące typy oddziaływań:

1. przypadki przyczynowości intencjonalnej, których nie da się sprowadzić do regularnego współwystępowania dwóch zdarzeń będących przejawem uniwersalnej korelacji,
2. przypadki opisywane przez teorię regularności,
3. przypadki oddziaływań pomiędzy ciałami w spoczynku (nie zdarzeniami), np. oddziaływanie grawitacyjne pomiędzy dwiema masami.

Każdy z powyższych typów oddziaływań ma swoją specyfikę, jednak najbardziej intrygujący jest typ pierwszy. Przypadki przyczynowości intencjonalnej są powszechnie

obecne i manifestują się poprzez procesy percepcji oraz tzw. działania inteligentne. Obydwa typy aktywności należą do podstawowych form intencjonalności. Formalnie – tak określoną przyczynowość – wyraża następujące zdanie warunkowe:

Jeśli *x* przyczynowo wywołuje *y*, to *x* łączy z *y* relacja przyczynowości intencjonalnej wtedy i tylko wtedy, gdy:

1. Albo (a) *x* jest stanem lub zdarzeniem intencjonalnym, a *y* warunkami spełniania *x* (lub ich częścią),
2. albo (b) *y* jest stanem lub zdarzeniem intencjonalnym, a *x* warunkami spełniania *y* (lub ich częścią),
3. jeśli zachodzi (a), to intencjonalna treść *x* wyjaśnia zajście *y* jako jego przyczynowo relewantny aspekt,
4. jeśli zachodzi (b), to intencjonalna treść *y* wyjaśnia zajście *x* jako jego przyczynowo relewantny aspekt (Searle, 1983, s. 122).

Relacja intencjonalnej przyczynowości istnieje wówczas, gdy stan intencjonalny S(p) (może on być albo przyczyną, albo skutkiem) jest jednym z członów relacji przyczynowej, a drugim jest stan świata (SŚ), określony przez warunki spełniania S(p). Gdy taki związek zaistnieje, twierdzi Searle, możemy mówić o tym, że albo S(p) spowodowało zajście SŚ (patrz: punkt 1), albo SŚ spowodował zajście S(p) (patrz: punkt 2). Określenie: „*X* spowodowało zajście *Y*" (*making something happen*) (Searle, 1983, s. 123) stanowi, zdaniem Searle'a, alternatywny – w odniesieniu do teorii regularności – sposób definiowania pojęcia przyczyny. „W najbardziej podstawowym sensie, kiedy C przyczynowo wywołuje E, C powoduje zajście E" (Searle, 1983, s. 123). Takie ujęcie nie przesądza – czy C regularnie będzie powodowało zajście E, czy okaże się, że jest to sytuacja jednorazowa. Regularność nie jest konstytutywna dla tak określonego pojęcia przyczyny.

Według Searle'a, w każdym przypadku, w którym przyczyną lub skutkiem jest stan intencjonalny, dany związek przyczynowo-skutkowy jest nam dany bezpośrednio i nie wymaga uprawomocnienia w regularności. Taki charakter związku przyczynowego potwierdzają nasze codzienne doświadczenia oraz hipotezy dotyczące pogłębiania rozumienia przyczynowości przez dzieci (objaśnię to w dalszej części rozdziału).

Dobrym przykładem prezentującym intencjonalny związek przyczynowo-skutkowy może być proces powstawania stanu percepcyjnego. Kiedy patrzymy na drzewo, to w naszym umyśle pojawia się stan, którego treść oraz warunki spełniania są zdeterminowane przez postrzegany obiekt. Znaczy to, że za powstanie stanu percepcyjnego odpowiada zrealizowany związek przyczynowy, w którym po stronie przyczyny znajduje się pewien obiekt, a po stronie skutku znajduje się stan intencjonalny o treści odnoszącej się do tegoż obiektu oraz związku przyczynowego, który doprowadził do jego zaistnienia (patrz: punkt 2. definicji), czyli wytworzony w umyśle percept wzrokowy. Analogiczną strukturę posiadają, zdaniem Searle'a, prior intencja, a także intencja w działaniu. Obydwa wymienione stany nie tylko reprezentują działanie lub prezentują ruch, ale również przyczynowo oddziałują na świat.

> *W każdym z [wymienionych] przypadków mamy do czynienia ze stanem lub zdarzeniem posiadającym samo-odniesienie przyczynowe o następującej formie: (w przypadku działań) częścią treści stanu lub zdarzenia intencjonalnego są takie warunki spełniania (w sensie: wymagań), które wymagają, by były one przyczyną «pozostałych» warunków spełniania (w sensie: wymaganych rzeczy) lub (w przypadku percepcji), to «pozostałe» warunki spełniania oddziałują przyczynowo na dany stan lub zdarzenie. […]. Kierunek dopasowania oraz kierunek przyczynowania są asymetryczne.*[28] (Searle, 1983, s. 122).

Searle poświęca najwięcej uwagi działaniom i percepcji, jednak równocześnie zaznacza, że zgodnie z zaproponowaną przez niego definicją wszystkie stany intencjonalne o niezerowym nakierowaniu na zgodność mogą wchodzić w relacje przyczynowe. Odnosi się to także do przywołanego wyżej przykładu o chęci napicia się wody.

Przytoczona definicja związku przyczynowego odnosi się nie tylko do struktury relacji przyczynowej, ale również do kwestii wyjaśniania. Użyte w definicji przyczynowości intencjonalnej określenie „relewantny aspekt" dotyczy wymagań, jakie musi spełnić **opis** konkretnego związku przyczynowego, by mógł on zostać uznany za poprawny. Jeśli dany opis nie uwzględni wystarczająco relewantnych aspektów zjawiska, wówczas nieuzasadnione będzie stwierdzenie: „x oddziałuje przyczynowo na y". Przykładowo, jeśli

[28] "In each case there is a self-referential Intentional state or event, and the form of the self-reference (in the case of action) is that it is part of the content of the Intentional state or event that its conditions of satisfaction (in the sense of requirement) require that it causes the rest of its conditions of satisfaction (in the sense of thing required) or (in the case of perception) that the rest of its conditions of satisfaction cause the state or event itself. […] In each case, cause and effect are related as Intentional presentation and conditions of satisfaction. Direction of fit and direction of causation is asymmetrical" (Searle, 1983, s. 122).

stwierdzenie kauzalne będzie miało postać: „to, co Sally zrobiła, wywołało zjawisko, które zauważył John" (Searle, 1983, s. 114), to trudno orzec, czy faktycznie działania Sally spowodowały zajście określonego związku przyczynowego. Jeśli jednak opis zjawiska uwzględni jego przyczynowo relewantne aspekty, wówczas nie powinno już być wątpliwości. W przytoczonym przykładzie taki opis mógłby wyglądać następująco: „Sally wstawiła wodę na gaz, a John zauważył, że woda się gotuje.". Fakt, że treść stanu intencjonalnego S(p) stanowi przyczynowo relewantny aspekt pewnego stanu świata, wskazuje, zdaniem Searle'a, na to, że pomiędzy S(p) a SŚ występuje logiczny związek. Nie jest on może tak silny, jak w przypadku definicji analitycznych typu: trójkąt to figura o 3 kątach, niemniej zależność pomiędzy S(p) a SŚ zdecydowanie nie jest przygodna.

Manipulowanie jako wzorzec rozpoznawania przyczynowości w świecie

Dotychczasowe rozważania dotyczyły związku przyczynowego, w którym jednym z członów relacji przyczynowej był stan intencjonalny, a drugim – stan świata. Związek taki istnieje zarówno wtedy, kiedy pierwszym członem jest stan intencjonalny, oddziaływujący na stan świata, jak i wtedy, kiedy pierwszym członem tej relacji jest stan świata, natomiast stan intencjonalny jest skutkiem jego oddziaływania. Wiemy jednak, że przeważająca większość związków przyczynowych to relacje zachodzące pomiędzy określonymi obiektami fizycznymi, które z intencjonalnością nie mają nic wspólnego, jak choćby w klasycznym przykładzie Hume'a z dwoma kulami bilardowymi (Hume, 1977, s. 31). Pojawia się zatem pytanie, jak można, wychodząc od przyczynowości intencjonalnej, uzasadnić przyczynowość czysto fizyczną, tzn. pozbawioną składnika intencjonalnego? Zdaniem Searle'a, przekonanie o występowaniu nieintencjonalnej relacji przyczynowej wywieść można z naszej zdolności do manipulowania różnego rodzaju przedmiotami oraz z występowania regularności towarzyszącej tego typu manipulacjom.

Aby wyjaśnić, jak wymienione elementy kształtują nasze przekonanie na temat przyczynowości w świecie fizycznym, amerykański filozof przytacza następujący przykład. Wyobraźmy sobie dziecko, które odkrywa, że można stłuc szklany wazon, rzucając w niego kamieniem. W takim przypadku, dziecko zauważa, że rezultatem intencji w działaniu związanej z wykonanym rzutem jest określony ruch ramienia, który prowadzi do przemieszczenia się kamienia i w konsekwencji – do rozbicia wazonu. Powtarzanie tego typu sekwencji w podobnych okolicznościach prowadzi do tego, że dziecko potrafi

zidentyfikować następujące etapy całego zjawiska: (1) „z pomocą" ręki, w której znajduje się kamień, można spowodować jego przemieszczenie się, (2) „za pomocą" poruszającego się z odpowiednią prędkością kamienia można rozbić wazon. Łatwo zauważyć, że relacja „za pomocą" jest przechodnia; jeśli „za pomocą" ruchu ręki można nadać kamieniowi odpowiednią prędkość oraz jeśli „za pomocą" poruszającego się z odpowiednią prędkością kamienia można stłuc wazon, to wynika z tego, że „za pomocą" ruchu ręki, w której znajduje się kamień, można stłuc wazon. Zdaniem Searle'a, przechodniość relacji „za pomocą" (*by-means-of*) powoduje, że przy odpowiedniej liczbie powtórzeń, poszczególne związki składające się na dany ciąg przyczynowy zostaną włączone w intencję w działaniu i staną się jej warunkami spełniania. W efekcie, intencja z postaci: „zobaczę, co się stanie, gdy rzucę z odpowiednią siłą kamień" – zmieni się w: „rozbiję wazon «za pomocą» operacji (1) i (2)". Co istotne, przechodniość relacji „za pomocą" zapewnia, że tak utworzoną intencję w działaniu cechuje tzw. samo-odniesienie przyczynowe, a to znaczy, że dostępne w treści intencji warunki spełniania wynikające z operacji(1), jak i (2) są przez nas doświadczane jako przyczyny zmian istniejących w świecie[29] (Searle, 1983, s. 128).

Przedstawiony przykład pokazuje również, że manipulowanie przedmiotami pozwala, metodą prób i błędów, odkryć regularności przyczynowe w świecie. Dziecko, realizując kolejne próby stłuczenia wazonu, zauważa z czasem, że w określonych przypadkach można, za pomocą odpowiednio ciężkiego kamienia, rozbić naczynie, jeśli jest ono zrobione z odpowiedniego materiału. Zdaniem Searle'a, zdolność dziecka do włączania w treść intencji zaobserwowanych nowych związków przyczynowych, pojawia się wraz ze zdolnością do odkrywania ich niezależności od własnych poczynań. W kontekście rozważanego przykładu jest to np. odkrycie, że twarde przedmioty, uderzające w naczynia, powodują ich zniszczenie, niezależnie od tego, czy zniszczenie nastąpiło w wyniku wykonanego rzutu kamieniem, czy być może na wazon spadł jakiś ciężki przedmiot ze stojącego obok regału. W ten sposób zamierzony ruch naszego ciała włączony zostaje w związek przyczynowy między nim a pewnym obiektywnym zdarzeniem, np. rozpadem naczynia na kawałki. Innymi słowy, naczynie nie rozpadłoby się, gdybym nie rzucił w nie kamieniem. Ponieważ potrafię powtarzać rzut kamieniem i nabieram w tym coraz większej wprawy, przeto regularność ta – rozpadnięcie się naczynia po tym, jak uderzy w nie

[29] Z czasem, kiedy dziecko nabierze wprawy w rozbijaniu naczyń poprzez rzucanie w nie kamieniem, wskazane działanie może uzyskać status działania podstawowego i wówczas poszczególne warunki spełniania, składające się na opisany ciąg przyczynowo-skutkowy, zostaną „skompresowane" do postaci działania typu „tłuczenie naczynia".

rzucony celowo kamień – może zostać wyodrębniona jako „coś więcej", niż zwykłe następstwo czasowe między zdarzeniami. To „coś więcej" odnosi się do mojego działania w wywołaniu zmiany w świecie. Znaczy to, że opanowanie umiejętności praktycznych, takich jak celny rzut kamieniem[30], wytwarzanie narzędzi i ich skuteczne użycie prowadzi w efekcie do wykształcenia się zdolności do rozpoznawania związków przyczynowych na podstawie wzorców wiążących działania praktyczne ze zmianami w świecie. Wzorce te są następnie ekstrapolowane na obiektywne relacje między zdarzeniami w świecie.

> *W przypadku, gdy [ktoś] obserwuje przyczynowość zdarzeń niezależnych od jego woli, nie doświadcza związku przyczynowego w taki sam sposób, w jaki doświadcza związku przyczynowego podczas działania lub percypowania, i pod tym względem zwolennicy Hume'a mają rację, twierdząc, że przyczynowość między niezależnymi od nas zdarzeniami nie jest obserwowalna w taki sam sposób, w jaki obserwowalne są same te zdarzenia. Lecz agent rzeczywiście obserwuje zdarzenia jako powiązane przyczynowo, a nie tylko jako sekwencję zdarzeń. To, że przypisuje on przyczynowość takiej sekwencji zdarzeń, jest uzasadnione lub może być uzasadnione, ponieważ to, co przypisuje w przypadku [takiej] obserwacji, jest tym, czego doświadczył w przypadku [własnej] manipulacji.*[31]

Przyczynowość intencjonalna a problem umysł-ciało

Trudno uciec od problemu umysł-ciało (*mind-body problem*), kiedy rozważa się zagadnienie przyczynowości intencjonalnej. Dla wielu neuronaukowców opowiedzenie się za istnieniem tego typu przyczynowości jest równoznaczne z akceptacją dualizmu, który współcześnie jest powszechnie kwestionowany nie tylko ze względu na filozoficzne trudności, które implikuje, ale głównie w związku z przytłaczającą liczbą danych empirycznych, których nie da się zinterpretować jako „wspierających" – niezależne od rzeczy fizycznych (*res extensa* w sensie Kartezjusza) – istnienie rzeczy umysłowych (*res cogitans* w sensie Kartezjusza) (Damasio, 2011; Koch, 2004). W związku z tym w

[30] Już Calvin wskazywał na to, jak ważną rolę odgrywało doskonalenie umiejętności celnego rzucania kamieniem oraz – jak umiejętność ta wpłynęła na pojawienie się mowy (Calvin, 1983).

[31] "In the case where he observes the causation of events independent of his will, he does not experience the causal nexus in the same way as he experiences the causal nexus in the experience of acting or perceiving, and in that respect the Humeans are right in claiming that causation between events independent of us is not observable in the way that the events themselves are observable. But the agent does observe the events as causally related, and not just as a sequence of events, and he is justified or can be justified in ascribing causality to such a sequence of events, for what he ascribes in the case of observation is something he has experienced in the case of manipulation." (Searle, 1983, s. 129).

neuronaukach świadomie unika się wyjaśnień odwołujących się do przyczynowości mentalnej. Polega to zwykle albo na próbie redukcji aspektu umysłowego do czynników lub zależności fizycznych, lub zastąpieniu pojęć odnoszących się do kategorii umysłowych przez pojęcia uznawane za „metafizycznie nieobciążone", takie jak nastawienie intencjonalne (*intentional stance*) (D. C. Dennett, 1997) lub program maszyny Turinga (Newell, 1990). W opinii Searle'a tego typu zabiegi nie są potrzebne, gdyż stany mentalne to takie same zjawiska biologiczne, jak laktacja, trawienie czy fotosynteza[32] (Searle, 1983, s. 264). Twierdzi on, że istnieje spójny opis relacji umysł-ciało, który nie musi prowadzić do dualizmu ani do aporii materializmu, np. do epifenomenalnego statusu stanów mentalnych (Searle, 1983, s. 255).

Zdaniem amerykańskiego filozofa, aby właściwie uchwycić status przyczynowości intencjonalnej w kontekście problemu umysł-ciało, należy wyjść od dość oczywistego spostrzeżenia: mózg to złożony system biologiczny. Składa się on, jak każdy tego typu twór, z szeregu prostych elementów (komórek nerwowych), których odpowiednie połączenia realizują złożone funkcje, co prowadzi do tego, że mózg przejawia, na poziomie systemu jako całości, zupełnie nowe cechy, których nie posiadają tworzące go elementy ani ich proste kombinacje. Istotne jest to, że stany umysłu są *przyczynowo wywoływane* przez odpowiednie operacje mózgu (*caused by*), a zarazem są one zrealizowane w jego strukturze (*realized*) (Searle, 1983, s. 265). Takie ujęcie, w opinii Searle'a, jest wolne zarówno od wątpliwości podnoszonych wobec podejścia dualistycznego (to, co mentalne, nie musi „przenikać" komórek nerwowych (patrz panpsychizm) ani na nie „odgórnie" oddziaływać (patrz koncepcja top-down causation), by uznać, że wpływa ono realnie na zachowania agenta), jak i fizykalistycznego (to, co mentalne, nie musi być ani epifenomenem, ani być identyczne z tym, co fizykalne). Najlepiej prześledzić sposób rozumowania Searle'a na podanym przez niego przykładzie płynności wody. Płynność to wysokopoziomowa cecha odpowiednio „zachowujących" się cząsteczek H_2O (inne ich „zachowanie" obserwuje się w stanie gazowym (para wodna), a jeszcze inne w przypadku stanu stałego (lód)). O żadnej z cząsteczek wody, twierdzi Searle, nie możemy powiedzieć, że jest ona płynna lub wilgotna, ale możemy stwierdzić, że (1) płynność **wywołana** jest przyczynowo przez zachowanie cząsteczek (*caused by*) oraz (2), że cecha ta jest **zrealizowana** (*realized*) w zbiorze cząsteczek H_2O. Pierwsze stwierdzenie wydaje się dość

[32] Searle określa swoje stanowisko mianem biologicznego naturalizmu (Searle, 1983, s. 264). W ten sposób dystansuje się od dominujących w filozofii umysłu takich stanowisk jak: funkcjonalizm, fizykalizm, dualizm czy behawioryzm.

oczywiste. Łatwo zauważyć, że jeśli wpłyniemy na zachowanie cząsteczek H_2O, to natychmiast zmieni się również ich makroskopowy stan skupienia. W zależności od prędkości poruszania się molekuł – mogą one przejść w stan płynny, gazowy lub stały. Dodatkowo, uzyskane przez zachowanie cząstek wysokopoziomowe cechy wody – same również funkcjonują przyczynowo. Kiedy woda jest w stanie płynnym, można ją przelewać pomiędzy naczyniami, można w niej coś wyprać lub się jej napić (por. Searle, 1983, s. 265). Wszystkie wymienione przypadki są możliwe z powodu płynności, za sprawą której woda, ujęta makroskopowo, sama oddziałuje przyczynowo na inne obiekty.

Searle przestrzega, by nie myśleć o płynności wody, jak o jakimś, „wydzielanym" przez cząsteczki H_2O „soku", dodatkowej substancji „ujawniającej się" w wyniku istniejących między molekułami interakcji. Kiedy opisujemy jakąś substancję jako płynną, to tak naprawdę opisujemy te same cząsteczki tylko na wyższym poziomie opisu, niż ten stosowany do pojedynczej cząstki. Płynność nie jest zatem epifenomenem, jest czymś jak najbardziej realnym, czymś **zrealizowanym** w molekularnej strukturze wody.

Z powyższego przykładu wnosić można, że zachowanie odpowiednio złożonego układu materialnego można w spójny sposób opisać zarówno na poziomie pojedynczych elementów (cząstek H_2O), jak i złożonego zbioru cząstek (płynności). O zjawisku opisanym na poziomie całego układu mówimy, że zostało ono **zrealizowane** za pomocą elementów niższego rzędu. Istotne jest również to, że o każdym z poziomów możemy orzec, że jest przyczynowo efektywny. Analogicznie, związane z działaniem intencjonalnym prior intencja lub intencja w działaniu – to wysokopoziomowe cechy złożonej sieci neuronalnej będącej w określonym stanie, który pojawił się w wyniku interakcji między neuronami wchodzącymi w jej skład. Stan intencjonalny, podobnie jak stan płynny w przypadku wody, można opisać na dwóch poziomach: na poziomie związków przyczynowych istniejących między poszczególnymi neuronami oraz na poziomie całej sieci neuronalnej. O wyższym poziomie możemy powiedzieć, że został zrealizowany za pomocą niższego. Istotne jest również to, iż stany intencjonalne mogą tworzyć związki przyczynowe z innymi stanami lub procesami umysłowymi, np. z reprezentacjami określonych zachowań. Omówioną analizę Searle'a można zobrazować za pomocą następującego schematu:

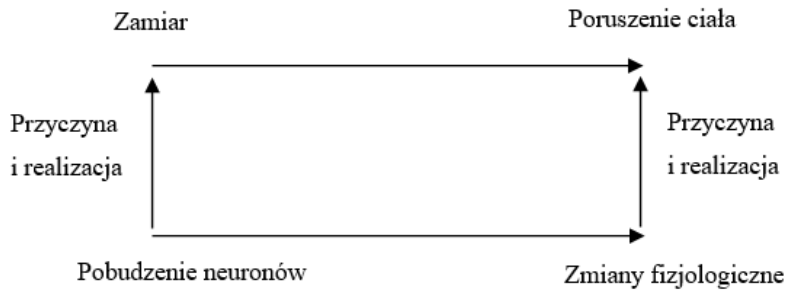


Diagram 4. Intencja w działaniu z perspektywy problemu umysł-ciało w ujęciu Searle'a.

W ujęciu wertykalnym schemat na Diagramie 4. wyraźnie wskazuje na istnienie dwóch rodzajów relacji funkcjonujących między niskopoziomowymi składnikami danego układu a jego wysokopoziomowymi cechami (przyczyna i realizacja). Natomiast w układzie horyzontalnym, w kolejnych odcinkach czasu – na każdym z poziomów opisu (mikro lub makro), pojawia się specyficzna dla danego poziomu relacja przyczynowa. Zdaniem Searle'a, takie ujęcie pozwala przezwyciężyć problemy metafizyczne, w które „wikła się" dualizm i fizykalizm.

> *Jeśli pomyślimy o związku między tym, co mentalne i tym, co cielesne jako o związku przyczynowym, pozostaniemy z tajemniczym pojęciem przyczynowości. Starałem się pokazać, że tak nie jest. Tak to odbieramy tylko wtedy, gdy myślimy o tym, co mentalne i o tym, co fizyczne jako o dwóch kategoriach ontologicznych, dwóch wzajemnie wykluczających się klasach rzeczy, rzeczach mentalnych i fizycznych, tak jakbyśmy żyli w dwóch światach, świecie mentalnym i świecie fizycznym. Jeśli jednak pomyślimy o sobie jako o istotach żyjących w jednym świecie, który zawiera rzeczy mentalne w takim samym sensie, w jakim zawiera rzeczy płynne i stałe, wówczas nie ma metafizycznych przeszkód, by tego typu rzeczy opisać w sposób przyczynowy. Moje przekonania i pragnienia, moja chęć napicia się czegoś oraz moje doświadczenia wzrokowe są rzeczywistymi stanami/cechami mojego mózgu, tak samo jak masywność stołu, na którym pracuję oraz płynność wody, którą piję, są cechami stołów i wody wywołane określonymi przyczynami.*[33]

[33] "If we think of the relation between the mental and the physical as causal, we are left with a mysterious notion of causation. I have argued that this is not so. It only seems so if we think of mental and physical as naming two ontological categories, two mutually exclusive classes of things, mental things and physical things, as if we lived in two worlds, a mental world and a physical world. But if we think of ourselves as living in one world which contains mental things in the sense in which it contains liquid things and solid things, then there are no metaphysical obstacles to a causal account of such things. My beliefs and desires,

Przedstawiona powyżej analiza pokazuje, że przyczynowość intencjonalną nie tylko da się włączyć do świata fizycznego, ale – jak argumentuje Searle – ma ona wręcz fundamentalny status. To z jej pomocą – poprzez manipulację – uczymy się odkrywać przyczynowość w świecie fizycznym. Z czasem potrafimy sprawnie identyfikować związki kauzalne, które są całkowicie od nas niezależne. Równocześnie, zdaniem Searle'a, uznanie istnienia przyczynowości intencjonalnej nie implikuje akceptacji stanowiska dualistycznego. Odpowiednie rozumienie, czym jest stan intencjonalny i w jakie relacje może wchodzić, pozwala uniknąć takiego zakwalifikowania.

Wybrane elementy omówionego wyżej stanowiska Searle'a wykorzystane będą w konstrukcji modelu złożonych działań intencjonalnych. Wprowadzona przez niego kategoria przyczynowości intencjonalnej nakazuje uznać „prior intencję", „intencję w działaniu" oraz inne stany intencjonalne za przyczynowo skuteczne, innymi słowy, za realnie wpływające na nasze zachowania. Bez naturalizującej interpretacji przyczynowości intencjonalnej poszczególne stany intencjonalne stałyby się szczególnego rodzaju fikcjami, pozbawionymi zdolności do wpływania na nasze zachowania. Nawet, jeśli w konkretnym przypadku działanie wywoływane jest przez intencjonalny stan umysłowy będący błędną reprezentacją (czego nie można wykluczyć), ogólna zasada, że prior intencja i intencja w działaniu są efektywne przyczynowo – nie zostaje, zdaniem Searle'a, naruszona.

Podsumowanie

Przeprowadzona przez amerykańskiego filozofa analiza działań intencjonalnych osadzona jest w szerszej teoretycznej strukturze – w teorii intencjonalności. Filozof ten, wbrew mnożącym się wątpliwościom, w zdecydowany sposób broni przyczynowego statusu intencji w działaniu oraz prior intencji. Co prawda, wymaga to rozszerzenia pojęcia intencjonalności, ale zabieg ten nie pociąga – przynajmniej w opinii jego autora – nadmiernych zobowiązań ontologicznych. Reasumując, powiedzieć można, że rdzeń koncepcji intencjonalności Searle'a tworzą trzy tezy: teza o nieeliminowalności intencjonalności, teza o jej biologicznej naturze oraz teza o przyczynowości intencjonalnej. Treść tej koncepcji omówiłem szczegółowo powyżej. Tu przypomnę te jej składniki, które wyrażone są we wspomnianych tezach.

my thirsts and visual experiences, are real causal features of my brain, as much as the solidity of the table I work at and the liquidity of the water I drink are causal features of tables and water." (Searle, 1983, s. 271).

Teza o nieeliminowalności intencjonalności głosi, że nie da się wyjaśnić ludzkiego działania bez uwzględnienia jego intencjonalnego charakteru, na który składają się: prior intencja, intencja w działaniu, których z kolei nie da się zrozumieć bez odniesienia do sieci stanów intencjonalnych. Zauważmy, że konsekwencją tej tezy jest odrzucenie postulatu redukcjonizmu rozumianego jako zastąpienie składników intencjonalnych i związków między nimi przez niskopoziomowe zależności fizyczne (w szerokim sensie słowa „fizyczny").

Teza o biologicznej naturze intencjonalności głosi, że stany intencjonalne mają biologiczny charakter i w tym sensie są częścią natury, posiadają pierwszoosobową ontologię, są zarazem wywoływane przez (*caused by*) neurobiologiczne procesy mózgowe, jak i w nich są zrealizowane (*realized*), ujawniając się jako wysokopoziomowe własności mózgu – tak ujęte stany umysłu są efektywne przyczynowo, tzn. realnie wpływają na nasze zachowania (Searle, 1992, 2008b).

Teza o przyczynowości intencjonalnej głosi, że intencjonalne stany (akty) umysłowe takie jak: pragnienia, intencje, przekonania – są przyczynowo powiązane z ludzkimi zachowaniami, a w konsekwencji – ze zmianami w świecie wywołanymi przez te zachowania. Intencjonalny stan (akt) umysłowy może być zarówno przyczyną (np. intencja w działaniu), a więc czynnikiem wywołującym zachowanie, jak i skutkiem, czyli stanem umysłowym, wywołanym przez określony stan (lub kompleks stanów) świata (np. stan percepcyjny). Podkreślić trzeba, że postulowana przez Searle'a forma przyczynowości nie pociąga – w jego opinii – konieczności wykroczenia poza zależności świata fizycznego. Ponieważ intencjonalność jest dla niego zjawiskiem biologicznym, a więc sama należy do świata fizycznego, przeto związek przyczynowy, którego jednym z członów jest akt intencjonalny, również należy do tego świata. Spośród powyższych tez najwięcej wątpliwości budzi ta ostatnia. Krytycy przyczynowości intencjonalnej m.in. Kim (Kim, 1995), Meijers (Meijers, 2000) argumentują, że propozycja Searle'a tylko z pozoru radzi sobie z kwestią sprawczości tych stanów. Utrzymują oni, iż uzasadnienia (którego Searle nie przedstawił) wymaga twierdzenie, że status intencjonalności przysługującej stanowi umysłowemu jest taki sam, jak status pojawiających się dopiero na poziomie makro własności posiadanych przez stany fizyczne. Zakładają przy tym, że ujęcie takie prowadzi do naddeterminacji przyczynowej (dany skutek wywoływany jest przez wiele przyczyn równocześnie, choć wystarczyłaby tylko jedna z nich) oraz do braku przyczynowego domknięcia świata fizycznego (to, co mentalne jest przyczyną tego, co fizyczne). Searle

szczegółowo odniósł się do wskazanych zarzutów (Searle, 1995; Searle, 2000). Nie wchodząc w szczegóły przywołanej tu polemiki całkowicie odrzucił on argumenty oponentów, wskazując, że posługują się oni zbyt uproszczonym pojęciem związku przyczynowego (Kim, Meijers) lub nie dość konsekwentnie odrzucają kartezjański sposób myślenia o tym, co mentalne i o tym, co fizyczne (Meijers).

> *Dlaczego ktokolwiek miałby wątpić w takie podejście? To dość standardowy materiał podręcznikowy. Dlaczego naddeterminacja wydaje się być problemem? Moja diagnoza błędu Meijersa i Kima jest taka, że zaakceptowali oni pewne standardowe błędy obecne w historii filozofii. Meijers uważa, że jeśli to, co «mentalne» i to, co «fizyczne» mają w ogóle odnosić się do czegoś rzeczywistego, to muszą należeć do różnych ontologicznie sfer (Kartezjusz), a ponadto zakłada on, że wszystkie związki przyczynowe istnieją wyłącznie między dyskretnymi zdarzeniami w czasie według schematu: zdarzenie wcześniejsze – przyczyna, jest warunkiem (wystarczającym? koniecznym? oboma?) zdarzenia późniejszego, czyli skutku (Hume). Nalegam, byśmy wreszcie zapomnieli o błędach naszych wybitnych przodków, w tym Arystotelesa i skupili się na tym, jak naprawdę działają systemy naturalne.*[34]

Z perspektywy celu niniejszej pracy omówienie i ocena filozoficznych polemik z tezami Searle'a mają znaczenie drugorzędne. Tym, co zdecydowało o wyborze jego koncepcji, była jej „moc inspirująca" w tworzeniu modelu działania intencjonalnego. Mam tu na uwadze przede wszystkim tezę o nieeliminowalności intencjonalności i pochodne względem niej charakterystyki prior intencji i intencji w działaniu. Traktowałem analizy Searle'a, konstruując zintegrowany model złożonych działań intencjonalnych, jako inspirację przy wyróżnieniu w proponowanym przeze mnie modelu dwóch ważnych podsystemów: podsystemu zarządzania siecią stanów intencjonalnych oraz podsystemu planowania i realizacji planów. Obydwa podsystemy stanowią bezpośrednie nawiązanie do teorii intencjonalności i w dużym stopniu respektują poniższe wnioski i spostrzeżenia Searle'a:

[34] "Why would anyone have any doubts about this account? It is fairly standard textbook stuff. Why does overdetermination seem as if it might be a problem? My diagnosis of Meijers's error and Kim's as well is that they have accepted certain standard errors from the history of philosophy. Meijers thinks that 'mental' and 'physical', if they name real phenomena, must name phenomena in different realms (Descartes), and he apparently thinks that all causation is a relation between discrete events in time whereby the earlier, the cause, is the condition (sufficient? necessary? both?) of the later, the effect (Hume). I am urging that we should forget about the mistakes of our distinguished forebears including Aristotle, by the way and describe how natural systems actually work." (Searle, 2000, s. 173).

- intencja to stan S(p), którego nakierowanie na zgodność następuje w kierunku świat→umysł; oznacza to, że treść stanu intencjonalnego będzie się do czegoś odnosić wówczas, gdy świat dostosuje się do umysłu,
- warunki spełniania składające się na treść intencji mają własność samoodniesienia przyczynowego, innymi słowy, część warunków spełniania wskazuje na to, co ma zostać w świecie zmienione, a część na to, by do takiej zmiany doprowadzić,
- działanie może zostać uznane za zaplanowane tylko wówczas, gdy wynikające z niego zachowanie będzie przyczynowo powiązane z prior intencją;
- treść intencji – zgodnie z zasadą holizmu – zawsze związana jest z innymi stanami intencjonalnymi należącymi do tzw. sieci,
- realizacja warunków spełniania intencji w przypadku działań podstawowych następuje przy wykorzystaniu określonej umiejętności należącej do dyspozycji tła,
- istnieją dwa typy intencji: (1) prior intencja oraz (2) intencja w działaniu; pierwsza pojawia się w związku z działaniami zaplanowanymi, a druga jest kluczowym elementem przeżycia towarzyszącego realizacji działań spontanicznych oraz zaplanowanych.

Wymienione tezy mają swoje odzwierciedlenie w projektowanym modelu. Wszystkie one poddane zostały konfrontacji z właściwą dla tej dziedziny, aktualną wiedzą teoretyczno-empiryczną. Wyniki badań przeprowadzonych przez psychologów intencji, w szczególności wnioski wyprowadzone z badań Libeta i Haggarda (omawiam je w rozdziale 4.) posłużyły do reinterpretacji Searle'owskiej intencji w działaniu oraz do wyróżnienia w jej obrębie kilku składowych, m.in.: chęci wykonania ruchu (*sense of urge*), odniesienia do docelowego obiektu lub zdarzenia (*reference forward to the goal object or event*), poczucia sprawstwa (*sense of agency*). Natomiast koncepcja tła została doprecyzowana za pomocą podsystemu uczenia się ze wzmacnianiem, dla którego bazą są dane i hipotezy z zakresu neurobiologicznych podstaw procesów decyzyjnych, które omawiam w następnym rozdziale.

3 Stany intencjonalne (idee) jako nagrody

Historia rozwoju człowieka to także historia zmian w sposobach jego działania. Pojawianiu się nowych typów narzędzi, a także nowych form organizacji społecznej towarzyszyło poszerzanie się klasy dostępnych człowiekowi celów oraz opanowywanie coraz bardziej złożonych działań intencjonalnych, podejmowanych w związku z chęcią ich osiągnięcia. Działania te obejmowały zwykle złożoną sekwencję zachowań, a ich planowanie i realizacja wymagały zaawansowanych kompetencji, wykraczających poza umiejętność wykonania kilku ruchów, wypowiedzenia paru słów czy wyciągnięcia prostych wniosków. Wydaje się, że to właśnie opanowanie takich złożonych, skonwencjonalizowanych sekwencji zachowań legło u podstaw interakcji społecznych, w których obie strony nauczyły się traktować cudze (ale także i własne) zachowania nie jako kompleksy cielesnych ruchów, lecz jako działania nasycone znaczeniem. Zaznaczyć trzeba, że klasa tego typu działań pojmowana jest szeroko. Należą do niej nie tylko wypowiedzi językowe, ale także różnego rodzaju zachowania traktowane jako nakierowane na osiągnięcie jakiegoś celu.

Kiedy obserwujemy czyjeś zachowanie i kwalifikujemy je jako celowe, to interpretując je nie możemy oderwać (w trybie abstrahowania) czysto biologicznego ruchu od celu, dla którego został on zainicjowany. Innymi słowy, znaczeniem takiego działania jest cel, który ma być poprzez nie urzeczywistniony. Dobrze obrazuje to czynność głosowania przez podniesienie ręki. W tym przypadku ruch ręki traktowany wyłącznie jako zmiana położenia jednej z kończyn pozbawiony jest jakiegokolwiek znaczenia. Zgodnie z podejściem tradycyjnym – do takiego ruchu cielesnego można dołączyć całą gamę znaczeń. Może on być rozumiany jako pozdrowienie, wskazanie czegoś, co dzieje się ponad głową, zgłoszenie chęci zabrania głosu czy wreszcie – zgodnie z przywołanym wyżej przykładem – jako zakomunikowanie swojego wyboru w głosowaniu. Za nietrafne uważam podejście, zgodnie z którym działanie znaczące powstaje na skutek „doklejenia” znaczenia do czysto

biologicznego ruchu. W mojej opinii, nie da się w przypadkach takich jak głosowanie, pozdrawianie, zgłaszanie się w celu zabrania głosu, itp. oddzielić znaczenia od zachowania, są one niejako „zrośnięte" ze sobą. Jeśli nawet w sytuacji głosowania czy zakomunikowania chęci zabrania głosu wydaje się, że mamy do czynienia z takim samym lub bardzo podobnym ruchem ręki, to jest to wrażenie złudne. W przypadku głosowania ruch wykonywany jest na komendę, jednocześnie przez wiele osób, a ręka trzymana jest w górze przez stosunkowo krótki czas i opuszczana na sygnał dany przez prowadzącego głosowanie. Natomiast zgłaszanie się do głosu przez podniesienie ręki jest autonomiczną decyzją każdego podmiotu z osobna, a czas, w trakcie którego trzymana jest w górze, zależy od szeregu okoliczności. Jak łatwo zauważyć, te pozornie niewielkie różnice w ruchu ręki zależą od reguł kulturowych (jedne określają procedurę głosowania, drugie – procedurę uczestniczenia w dyskusji i zabierania w niej głosu), a więc to składniki znaczenia określają ostateczną formę ruchu ręki. Takich subtelnych różnic w pozornie tożsamych cielesnych ruchach jest zresztą więcej i są one związane z tym, iż każde z owych działań uzyskuje znaczenie ze względu na grę społeczną, w której jest usytuowane. Określone ruchy nabierają konkretnego znaczenia dopiero w kontekście **gry**, w której zostają wykonane, czyli użyte. Sygnalizuję tę kwestię, by pokazać, jak złożona jest natura działania znaczącego i jak daleko nam do pełnego jej zrozumienia.

Na ogół tego typu działaniom towarzyszą również złożone intencje, składające się z szeregu warunków spełniania, a co za tym idzie – ze złożonej treści dostępnej w formie wielopoziomowego opisu. Przedstawiony w rozdziale drugim przykład zabójstwa arcyksięcia Franciszka Ferdynanda pokazuje, że pojedyncze działanie może być wpisane w liczne i istotnie różne konteksty. Wymienione cechy działań intencjonalnych nie pojawiają się spontanicznie, zwykle są one efektem procesu uczenia się realizowanego za pomocą odpowiednio zaawansowanego mechanizmu. Tego typu mechanizm jest skuteczny, jeśli umożliwia: (1) nabywanie nowych zdolności, (2) dostosowywanie już nabytych umiejętności do zmieniających się warunków, (3) łączenie poszczególnych zdolności w większe całości, bez konieczności ponownego uczenia się. Na pierwszy rzut oka wydaje się, że spełnienie wymienionych warunków pociąga wysoki stopień skomplikowania takiego mechanizmu. Tymczasem badania z obszaru uczenia maszynowego, w szczególności badania dotyczące tzw. uczenia się ze wzmacnianiem, pokazują, że odpowiednio zaprojektowana struktura, wykorzystująca metodę prób i

błędów, pozwala – przy pewnych dodatkowych założeniach – spełnić, w dużym zakresie, wszystkie powyższe wymagania.

W tym kontekście pojawia się pytanie: jak tego typu algorytm – wraz z jego funkcjonalnymi własnościami – można odnieść do pracy ludzkiego mózgu/umysłu? Odpowiedź na nie pojawiła się w latach 90-tych XX wieku. Zespół z Salk Institute (Schultz i in., 1997) zaproponował wówczas, na podstawie badań nad procesami uczenia się warunkowego makaków, tzw. hipotezę dopaminergicznego błędu predykcji nagrody (HDBPN). Zgodnie z tą hipotezą, fluktuacje wyładowań neuronów dopaminergicznych można zrekonstruować za pomocą tzw. algorytmu TDRL, jednej z metod uczenia się ze wzmacnianiem. Od wielu lat prowadzone są badania weryfikujące wiarygodność tej hipotezy. Wielu badaczy próbuje określić jej zasięg oraz osadzić ją w szerszym kontekście ludzkich zachowań. Read Montague w pracy *Why choose this book. How we make decisions?* zaproponował ciekawą syntezę dotychczasowych wyników związanych z HDBPN (Montague, 2006). Zaproponowana przez niego interpretacja danych uzyskanych w wielu badaniach pozwala lepiej zrozumieć neurobiologiczne „zaplecze" ludzkich działań, a także – co szczególnie ważne – pokazuje, jak można wyjaśnić przynajmniej niektóre złożone zachowania za pomocą stosunkowo prostej aparatury pojęciowej. Jest to istotne z perspektywy niniejszej pracy, gdyż wykorzystane przez Montague metody obliczeniowe pozwalają zdefiniować relację pomiędzy określonymi stanami intencjonalnymi a zachowaniami. Wyniki te wprost odnoszą się do głównego celu doktoratu, którym jest opracowanie modelu złożonych działań intencjonalnych.

Badania nad HDBPN liczą sobie zaledwie kilkanaście lat, dlatego wiele zagadnień związanych z tą hipotezą nadal czeka na weryfikację eksperymentalną (Colombo, 2014). Uważam jednak, że pomimo tych ograniczeń, warto przeprowadzić dyskusję teoretyczną, aby oszacować, w jakim stopniu opracowany w neurobiologii model selekcji zachowań może zostać wykorzystany w badaniach nad działaniami intencjonalnymi. Nie jest to zadanie proste, gdyż z jednej strony mamy do czynienia z modelem, który bezpośrednio odnosi się do problemu kontroli zachowań i dostarcza szeregu niebanalnych hipotez w tej kwestii. Z drugiej jednak strony, aparat pojęciowy wykorzystany w badaniach neurobiologicznych dotyczących HDBPN jest zupełnie inny, niż ten stosowany w psychologii czy filozofii. W niniejszym rozdziale przedstawiam rozwiązania pozwalające zniwelować zasygnalizowane różnice pojęciowe. Tym samym pokazuję, w jaki sposób

wnioski płynące z hipotezy HDBPN można włączyć w projekt konstrukcji zintegrowanego modelu działań intencjonalnych.

W pierwszej kolejności omówię nieco dokładniej hipotezę dopaminergicznego błędu predykcji nagrody, wskażę też zjawiska i odkrycia, które doprowadziły do jej sformułowania. W kolejnym punkcie zaprezentuję model obliczeniowy hipotezy oraz jego najważniejsze cechy. Następnie odniosę się do tego, na ile zaprezentowany model daje się zastosować w badaniach nad działaniami intencjonalnymi. W ostatniej części rozdziału omówię te metody uczenia się ze wzmacnianiem, które – spośród dostępnych rozszerzeń – mogą potencjalnie wspomóc proces konstrukcji zintegrowanego modelu działań intencjonalnych. W ten sposób przygotowany zostanie grunt pod końcowy rozdział dysertacji, w którym mechanizm uczenia się ze wzmacnianiem stanie się jednym z kluczowych podsystemów ZMDI.

3.1 Hipoteza dopaminergicznego błędu predykcji nagrody (HDBPN)

„Najbardziej ekscytująca fraza w nauce, ta która oznajmia nowe odkrycia, to nie «Eureka» (znalazłem), ale «to zabawne»", twierdzi Isaac Asimov (Montague, 2006, s. 108). Trudno ocenić, na ile tego typu stwierdzenie faktycznie towarzyszy wszystkim ważnym odkryciom naukowym. Wiemy jednak, z relacji Reada Montague, że dwukrotnie było ono wypowiedziane podczas badań nad dopaminą – neurotransmiterem wpływającym m.in. na przebieg działania procesów motywacyjnych, poznawczych i motorycznych, zarówno u ludzi, jak i u zwierząt, u których występuje ośrodkowy układ nerwowy. Pierwsze obserwacje dotyczące roli dopaminy datuje się na połowę lat 50-tych XX wieku (Marsden, 2006), kiedy w badaniu mózgów zmarłych osób, chorujących wcześniej na chorobę Parkinsona, zidentyfikowano bardzo niski poziom dopaminy. W kolejnych latach udało się rozpoznać wpływ dopaminy na przebieg schizofrenii (tzw. hipoteza dopaminowa), na różnego rodzaju uzależnienia oraz na związane z układem nagrody mechanizmy uczenia się. Z czasem zidentyfikowane zostały poszczególne szlaki dopaminergiczne (szlak mezolimbiczny, szlak mezokortykalny, szlak nigrostriatalny, szlak tuberoinfundibularny) oraz kluczowe struktury, w których dochodzi do produkcji dopaminy (brzuszna część pola nakrywki, istota czarna) (Stahl, 2009, s. 23–31). Choć badania nad funkcją dopaminy prowadzone są od ponad 60 lat, to nadal istnieje poważna trudność w rozpoznaniu wspólnej zasady, która wyjaśniałaby poszczególne przypadki chorobowe oraz określone typy

zachowania. Hipoteza dopaminergicznego błędu predykcji nagrody, sformułowana przez zespół badaczy z Salk Institute w 1991 roku, może się okazać pomocna w przezwyciężeniu trudności, o której mowa powyżej. Do stworzenia tej hipotezy przyczyniły się przede wszystkim badania Wolframa Schultza, dotyczące fluktuacji wyładowań neuronów dopaminergicznych w kontekście procesu uczenia się ze wzmacnianiem. Schultz, zanim jeszcze został zidentyfikowany związek dopaminy z uczeniem się, podobnie jak wielu innych naukowców, rozpoczął swoje badania od opisania związku dopaminy z motoryką, która – jak wiadomo – zostaje radykalnie zaburzona w chorobie Parkinsona. Występujący w chorobie zanik aktywności neuronów dopaminergicznych powoduje sztywność, trudność w inicjowaniu ruchów oraz ich spowolnienie. Badacze skupili swą uwagę głównie na relacji pomiędzy zdolnościami motorycznymi a stężeniem dopaminy w określonych strukturach mózgowych (Birkmayer & Hornykiewicz, 1962). Skuteczność leku L-Dopa, opracowanego w latach 70-tych XX wieku, zwiększającego poziom dopaminy w mózgowiu, potwierdziła, że w sposób zasadniczy dopamina wpływa na dyspozycje motoryczne pacjentów. Niejako równolegle do wskazanych badań prowadzone były również badania nad zwierzętami (szczurami i małpami) realizowane w paradygmacie BSR (*Brain Stimulation Reward*)[35]. Realizowane eksperymenty polegały na stymulacji szlaku mezolimbicznego, w ramach którego neurony dopaminergiczne pełnią kluczową rolę. Potwierdziły one, że dopamina jest odpowiedzialna za efekt wzmocnienia, który – wielokrotnie powtórzony – prowadzi do zachowań kompulsywnych, charakterystycznych dla stanu uzależnienia (Wise, 2002).

Wiedza o motorycznej i nagradzającej funkcji dopaminy zainspirowała Wolframa Schultza do zainicjowania badań nad uczeniem warunkowym małp, w którym współwystępują obydwie funkcje. Schultz zauważył, mierząc aktywność neuronów dopaminergicznych, że wyróżnić można trzy charakterystyczne wzorce wyładowań. Pierwszy wzorzec to niemal jednorodny poziom wyładowań, który był interpretowany przez wielu badaczy zajmujących się chorobą Parkinsona jako tryb bezczynności (*idle*).

[35] Paradygmat *Brain stimulation reward* opracowany został przez Jamesa Oldsa oraz Petera Milnera w 1953 roku. Podczas eksperymentów na szczurach, które polegały na elektrycznej stymulacji wybranych struktur mózgowych, zauważyli oni, że zwierzęta preferują te miejsca w otoczeniu, w których doszło do stymulacji jądra przegrody. Uznali, że wskazany efekt można wyjaśnić, przyjmując, że struktura ta odpowiada za reprezentowanie wzmocnień, innymi słowy, określone miejsca zyskały swój wyróżniony status, ponieważ skojarzone zostały ze stanem nagrody. Potwierdzeniem przedstawionego wniosku okazały się dalsze eksperymenty. W jednym z nich udało się nauczyć szczury zachowania polegającego na naciskaniu dźwigni generującej ciąg impulsów elektrycznych stymulujących jądro przegrody. Z czasem tego typu stymulacja zaczęła być używana jako wzmocnienie instrumentalne (*operant reinforcer*), pozwalające kształtować szeroki wachlarz zachowań (Olds & Milner, 1954).

Drugi wzorzec to wzrost wyładowań neuronów w momencie otrzymania nagrody (np. porcji soku) przez małpę. Trzeci wzorzec charakteryzuje się znaczącym obniżeniem aktywności neuronów dopaminergicznych w chwili, gdy zwierzę spodziewa się otrzymania nagrody, ale jej nie uzyskuje ze względu na popełniony błąd lub manipulację eksperymentatora (Schultz i in., 1993). Zauważono również, że wzorzec drugi pojawia się także wtedy, gdy pozyskanie nagrody skojarzone zostanie ze stale poprzedzającym ją bodźcem – np. z dźwiękiem drzwiczek od skrzynki zawierającej tę właśnie nagrodę. Podobny efekt można uzyskać włączając do procedury uczenia kolejne, całkowicie niezwiązane z nagrodą bodźce, jak chociażby błysk światła czy dźwięk dzwonka. Ciekawe jest również to, że wzrosty wyładowań skojarzone z momentem podania nagrody zaczynają stopniowo zanikać wraz z kolejnymi powtórzeniami. W rezultacie – po nauczeniu się przez zwierzę sekwencji prowadzącej do uzyskania nagrody – pozostaje jedynie wzrost związany z pierwszym bodźcem sygnalizującym pojawienie się nagradzającej sekwencji. Zgromadzone dane świadczyły, zdaniem Schultza, o istotnym związku dopaminy z procesem uczenia się, w tym – z identyfikacją bodźców ważnych z perspektywy realizowanego zadania.

W modelu obliczeniowym, opracowanym przez Rescorla i Wagnera, który dotychczas był stosowany do uczenia się ze wzmacnianiem, niestety, nie uwzględniono wszystkich zaobserwowanych przez niemieckich badaczy zjawisk (model ten nie był przydatny głównie w odniesieniu do efektu blokowania obserwowanego w sytuacjach, gdy dany bodziec warunkowy zaczyna być poprzedzany nowym bodźcem warunkowym (Montague, 2006, s. 108)). Przełom nastąpił, kiedy w 1993 roku doszło w Salk Institute do współpracy Wolframa Schultza z Peterem Dayenem, młodym matematykiem zajmującym się metodami uczenia maszynowego. Szczególnie cenna okazała się wiedza Dayena dotycząca algorytmów uczenia się ze wzmacnianiem, które od początku lat 80-tych ubiegłego wieku zaczęły się dynamicznie rozwijać. Dayen zauważył, że zidentyfikowane przez Schultza wzorce wyładowań neuronów dopaminergicznych i rejestrowane na tej podstawie czasowe fluktuacje w aktywności neuronów doskonale pasują do tzw. błędu predykcji nagrody (TDRL, *temporal difference reinforcement learning*), zdefiniowanego w algorytmie uczenia się ze wzmacnianiem opartego na różnicach czasowych. Odkrycie związku pomiędzy błędem predykcji nagrody a wzorcami wyładowań neuronów dopaminergicznych po raz pierwszy pozwoliło w pełni wyjaśnić rolę dopaminy w procesie uczenia się. W nowym kontekście wszystkie trzy wzorce wyładowań neuronów

dopaminergicznych (wymienione wyżej) zyskały swoją precyzyjną interpretację, odwołującą się do działania algorytmu TDRL (więcej na ten temat w dalszej części rozdziału).

Obecnie hipoteza dopaminergicznego błędu predykcji nagrody (HDBPN) jest weryfikowana głównie w ramach badań neuroekonomicznych. Według mojej wiedzy, do tej pory nie sformułowano następującego pytania: czy można, a jeśli tak, to w jaki sposób, odnieść HDBPN do działań intencjonalnych? Z perspektywy niniejszej pracy jest to pytanie zasadnicze: jak zastosować zidentyfikowany przez Schultza i Dayena model uczenia się ze wzmacnianiem do wyjaśnienia działań intencjonalnych? Aby na nie odpowiedzieć, należy najpierw zrekonstruować główną zasadę działania algorytmu RL (*reinforcement learning*) oraz jego najważniejsze cechy.

3.2 Algorytm RL jako model obliczeniowy HDBPN

Zasada działania algorytmu

W 1954 roku Marvin Minsky, pionier badań nad sztuczną inteligencją, zainspirowany badaniami odnoszącymi się do uczenia warunkowego zwierząt, opracował pierwszy algorytm uczenia się ze wzmacnianiem (RL, *reinforcement learning*), inicjując w ten sposób powstanie nowej klasy algorytmów w obszarze uczenia maszynowego. Przez wiele lat sądzono, że metoda zaproponowana przez Minsky'ego nie sprawdza się w praktycznych zastosowaniach, gdyż proces uczenia się zachowań celowych wyłącznie na podstawie informacji wartościujących, pozyskiwanych w trakcie interakcji agenta ze środowiskiem, jest zbyt powolny. Ta niekorzystna opinia zmieniła się w latach 80-tych XX wieku. Sutton i Barto (por. Sutton, 1998) wykazali podówczas, że algorytm Minskiego może być skutecznym narzędziem służącym do rozwiązywania określonych problemów, gdy zostanie uzupełniony o dodatkowe założenia dotyczące środowiska (Montague, 2006, s. 92).

Ogólną ideę algorytmu uczenia się ze wzmacnianiem można przedstawić w postaci następującego diagramu:

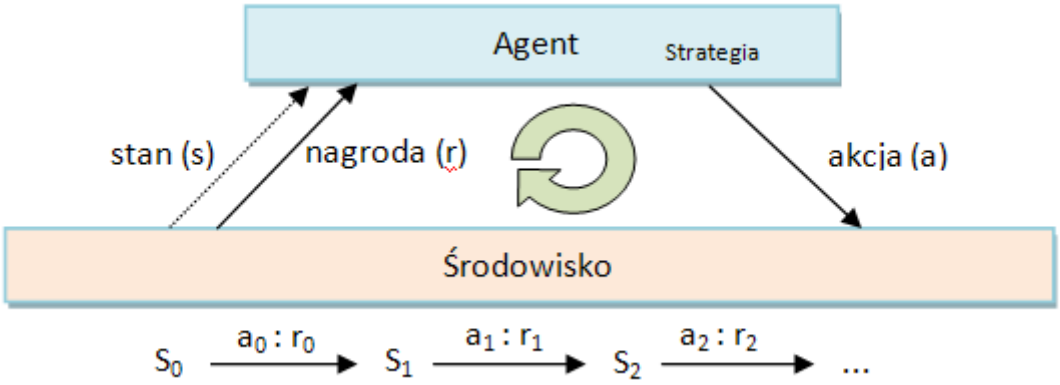


Diagram 5. Schemat algorytmu uczenia ze wzmacnianiem (za: Van Dijk 2008).

Działanie algorytmu rozpoczyna się w stanie startowym S_0. Następnie agent wykonuje pierwszą akcję a_0[36], która powoduje przejście do stanu S_1 oraz uzyskanie nagrody r_0. Cykl ten jest powtarzany w kolejnych odstępach czasu, a w zależności od przyjętej strategii, czyli reguły wyboru akcji, może prowadzić do pełnej lub częściowej eksploracji środowiska. Pełna eksploracja oznacza, że agent odwiedzi wszystkie stany środowiska, innymi słowy będzie posiadał kompletną wiedzę o jego strukturze, co pozwoli mu z czasem znaleźć strategię optymalną. Natomiast w przypadku częściowej eksploracji pewna grupa stanów świata nigdy nie zostanie przez agenta poznana, co w konsekwencji może prowadzić do pomijania miejsc zawierających atrakcyjne nagrody. Taka sytuacja może oznaczać, że wypracowana przez agenta strategia będzie co najwyżej suboptymalna. Zakłada się przy tym, że agent nie musi dysponować żadnym apriorycznym modelem środowiska (*model free*) w punkcie wyjścia, potrafi jedynie: wykonywać akcje, rozpoznawać nagrody oraz – na podstawie obserwacji – identyfikować określone stany świata. W początkowej fazie wybór akcji odbywa się na podstawie arbitralnie przyjętego sposobu działania, np. ma charakter losowy. Od agenta wchodzącego w interakcje ze środowiskiem oczekuje się, że nauczy się on takiej strategii (*policy*) wybierania zachowań, która prowadzić będzie do „maksymalizacji nagród długoterminowo" – bez względu na to, od jakiego stanu środowiska agent rozpocznie swoje działanie (jest to tzw. uczenie się na

[36] Tłumaczę angielskie słowo *action* jako „akcja", a nie „czynność" czy „działanie". Czynię tak, aby zachować zgodność z terminologią stosowaną w literaturze o uczeniu maszynowym, por. np. przywoływane tu cytaty z: Cichosz, P. (2007). *Systemy uczące się* (Wyd. 2). Warszawa: Wydawnictwa Naukowo-Techniczne. Przyjmuję ponadto, że słowa „agent" oraz „akcja" odniesione do ludzi znaczą „podmiot" oraz „zachowanie".

podstawie opóźnionych nagród[37] (P. Cichosz, 2007, s. 717)). Należy zaznaczyć, że przebieg interakcji nie musi być deterministyczny i stacjonarny, tzn. wartość uzyskiwanych nagród może być realizacją zmiennej losowej, a osiągane stany środowiska po wykonaniu takiej samej akcji mogą być różne. Kluczowym składnikiem metod uczenia się ze wzmacnianiem jest tzw. funkcja wartości definiowana następująco:

gdzie: r(t) to realizacja zmiennej losowej reprezentującej wartość nagrody dla chwili t, przy założeniu, że agent znajduje się w stanie s, γ to współczynnik dyskonta ($\gamma \leq 1$) powodujący, że ta sama nagroda otrzymywana z opóźnieniem jest dla agenta mniej wartościowa, niż nagroda otrzymana wcześniej (P. Cichosz, 2007, s. 718). Funkcja $V^{\pi}(s)$ określa dla danego stanu s oraz pewnej strategii wybierania akcji π oczekiwaną zdyskontowaną sumę przyszłych nagród. Zakładając, że modelem środowiska, w którym funkcjonuje agent, jest tzw. proces decyzyjny Markowa (P. Cichosz, 2007, s. 727), problem znalezienia optymalnej strategii można zdefiniować jako problem decyzyjny Markowa. Można, korzystając z wymienionych założeń oraz równań Bellmana, określić funkcję $V^{\pi}(s)$ następująco: $\mathbf{V^{\pi}(s_t) = r_0 + \gamma V^{\pi}(s_{t+1})}$. Zapis ten oznacza, że

> *[...] dla pewnej ustalonej strategii przy obliczaniu funkcji wartości nie musimy rozpatrywać oczekiwanej zdyskontowanej sumy przyszłych nagród z nieskończonej liczby kroków czasu, a możemy ograniczyć się do oczekiwanych efektów wykonania*
>
> $$V^{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^{t} r_t\right]$$
>
> *jednej tylko akcji. Wystarczy wówczas wziąć pod uwagę oczekiwaną wartość nagrody oraz stany [środowiska], do jakich można trafić po jej wykonaniu. Wartości tych stanów zawierają bowiem pełną informację o tym, jakich nagród możemy oczekiwać w następnych krokach, a odpowiednio ważąc je prawdopodobieństwami przejść, otrzymujemy oczekiwaną wartość następnego stanu.* (P. Cichosz, 2007, s. 741).

Z wymienionej możliwości korzysta jeden z najbardziej popularnych algorytmów uczenia ze wzmacnianiem – metoda różnic czasowych (TD), wykorzystywana do rozwiązywania wieloetapowych problemów predykcyjnych.

[37] „Dobra strategia nie od razu musi przynieść dobre efekty, ale sprawdzi się w dłuższym horyzoncie czasowym." (P. Cichosz, 2007, s. 717).

> *W takich problemach należy na każdym etapie wygenerować prognozę pewnej nieznanej końcowej wartości na podstawie dostępnej w tym kroku cząstkowej informacji. Można przyjąć, że w kolejnych krokach informacja ta jest coraz bardziej pełna i wiarygodna, powinna więc umożliwiać coraz lepsze stawianie prognozy. W trakcie uczenia się predykcje generowane w poszczególnych krokach modyfikuje się za pomocą błędów (Δ) obliczanych jako różnice wartości przewidywanych w dwóch kolejnych krokach czasu, w jednym [s_t], którego dotyczy modyfikacja, oraz następnym [s_{t+1}], w którym prognoza przez domniemanie powinna być lepsza. [...] W przypadku wartościowania strategii dla celów uczenia się ze wzmocnieniem wartością «przewidywaną» w kroku t jest zdyskontowana suma przyszłych nagród, reprezentująca wartość stanu [st].* (P. Cichosz, 2007, s. 754).

Przedstawioną metodę definiują formalnie następujące kroki algorytmu TDRL:

1. *dla wszystkich kroków t wykonaj;*
2. *obserwuj aktualny stan* s_t*;*
3. *wybierz akcję* $a_t := \pi(s_t)$ *do wykonania w stanie* s_t*;*
4. *wykonaj akcje* a_t*;*
5. *obserwuj wzmocnienie* r_t *i następny stan* s_{t+1}*;*
6. $\Delta = r_0 + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t)$*;*
7. *uaktualnij*$^{\beta}$ $(V(s_t), \Delta)$*;*
8. *koniec dla.*[38] *(P. Cichosz, 2007, s. 755).*

W momencie zainicjowania algorytmu funkcja wartości dla wszystkich stanów świata może mieć przypisaną tę samą wartość np.: V(s)=0 (wartość początkowa V(s) teoretycznie nie ma wpływu na końcowe rozwiązanie, może jednak w znaczący sposób skrócić czas jego poszukiwania). W kolejnych chwilach algorytm – zgodnie z bieżącą strategią – realizuje określone akcje, powodując przejście agenta do kolejnych stanów środowiska. Na wczesnym etapie działania algorytmu strategia łączy w sobie dwa elementy: eksploracyjny (oparty w głównej mierze na losowym wyborze akcji) oraz zachłanny (wybierana jest ta akcja, która spowoduje przejście do stanu o największej wartości V – jest to tzw. eksploatacja). Odpowiednio zrównoważone eksploracja i eksploatacja zapewniają, że z

[38] Wyrażenie „koniec dla" pełni w tym przypadku funkcję komentarza określającego miejsce zakończenia pętli (dla języków programowania jest to typowa konwencja zwiększająca czytelność zapisu algorytmu).

jednej strony agent rozpozna wszystkie ważne, dostępne dla niego stany środowiska, a z drugiej – w pewnym momencie zacznie wykorzystywać zdobyte doświadczania do podejmowania optymalnych decyzji (wyboru akcji, które prowadzą do maksymalizacji sumy przyszłych nagród w dłuższej perspektywie).

Aby zrozumieć, jakie wartości może przyjąć błąd predykcji nagrody, warto rozważyć następujący przykład. Wyobraźmy sobie, że robot ma pozyskać nagrodę znajdującą się w krótkim korytarzu (patrz: rysunek poniżej). W takim przypadku korytarz pełni rolę środowiska, robot rolę agenta. Poszczególne miejsca w korytarzu, do których agent może trafić, to z kolei określone stany środowiska (oznaczone na rysunku s1 – s7). Ponadto zakłada się, że robot dysponuje dwoma rodzajami akcji: ruchem do przodu (akcja 1) lub ruchem do tyłu (akcja 2). Niezerowa nagroda znajduje się tylko w stanie s5 (patrz: czerwony prostokąt na Rys. 1). Przyjmuje się, że początkowo funkcja wartości V dla każdego stanu środowiska wynosi 0 (por. Rys. 2), a strategia, czyli reguła wyboru akcji, zdefiniowana jest następująco: zawsze próbuj kontynuować ruch zgodny z ruchem poprzednim, jeśli się to nie uda, zmień kierunek poruszania. Jeśli agent zacznie się poruszać do przodu, to tego typu ruch będzie kontynuowany do chwili, gdy dotrze on do nagrody w stanie s5. Stan s5 jest w tym przykładzie traktowany jako stan końcowy, który zamyka dany epizod. Po zakończeniu epizodu agent wraca do jednego ze stanów początkowych, czyli s1 lub s7. Wizualnie można tego typu środowisko przedstawić następująco:

s1	s2	s3	s4	s5	s6	s7
0	0	0	0	1	0	0

Rys. 1 Rozkład nagród w przykładowym środowisku robota.

s1	s2	s3	s4	s5	s6	s7
0	0	0	0	0	0	0

Rys. 2 Początkowy stan funkcji wartości V.

Przyjmijmy, że robot rozpocznie swoją eksplorację od stanu s1. Zgodnie z przyjętą polityką – kolejne akcje będą powodowały ruch robota do przodu lub do tyłu. Wykonanie akcji 1 w chwili t=0 dla stanu s1 skutkuje przejściem ze stanu s1 (kolor żółty – por. Rys.

3) do s2 (kolor zielony – por. Rys. 3) oraz pozyskaniem nagrody r=0. Na podstawie uzyskanych informacji agent może obliczyć błąd predykcji nagrody za pomocą następującej formuły:

- $\Delta = r_0 + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t) = 0 + 0{,}8*0 - 0 = 0$
- $V^{\pi}(s_t) = V^{\pi}(s_t) + \beta*\Delta = 0 + 0{,}5*0 = 0 = V^{\pi}(s_1)$

Stała 0.8, która występuje w powyższym wzorze, to tzw. współczynnik dyskonta (γ) decydujący o tym, na ile wartość nagród zostanie obniżona ze względu na ich opóźnienie w czasie. Jeśli $\gamma=0$, to agent skupia się wyłącznie na natychmiastowych nagrodach. Jeśli natomiast $\gamma=1$, to nagrody odroczone w czasie są dla agenta równie ważne, jak nagrody natychmiastowe. Z kolei wartość $\beta=0{,}5$ odnosi się do tzw. rozmiaru kroku. Z jej pomocą określa się wpływ danej aktualizacji na docelową wartość funkcji wartości. Jeśli dany problem wymaga, by wpływ kolejnych aktualizacji malał z czasem, wówczas stałą β zastępuje się odpowiednią funkcją zależną od czasu lub od wykonanych aktualizacji.

s1	s2	s3	s4	s5	s6	s7
0	0	0	0	0	0	0

Rys. 3 Stan środowiska po przejściu ze stanu s1 do s2.

W kolejnych chwilach robot będzie przemieszczał się do przodu, aż osiągnie stan s5 (lub do tyłu, jeśli eksploracja rozpocznie się od stanu s7). Po wykonaniu pierwszej iteracji funkcja wartości V dla poszczególnych stanów przyjmie postać:

s1	s2	s3	s4	s5	s6	s7
0	0	0	0,5	0	0,5	0

Rys. 4 Stan środowiska po przejściu całego korytarza - aż do stanu s7.

Osiągnięcie stanu s5 nie tylko kończy dany epizod, ale również powoduje, że robot zostaje przeniesiony do jednego ze stanów początkowych: s1 albo s7. Jeśli robot zostanie przeniesiony do stanu s7, to funkcja wartości – po zrealizowaniu kolejnego epizodu – przyjmie następującą postać:

s1	s2	s3	s4	s5	s6	s7

0	0	0	0,5	0	0,5	0

Rys. 5 Stan środowiska po cofnięciu się robota ze stanu s7 do s1.

Po 10 epizodach – przejściach korytarzem do przodu lub w tył – funkcja wartości dla poszczególnych stanów przyjmie następującą postać:

s1	s2	s3	s4	s5	s6	s7
0.38	0.58	0.78	0.99	0	0.99	0.40

Rys. 6 Wartości V dla stanów środowiska po wielokrotnym przejściu korytarza.

Łatwo zauważyć, analizując wartości poszczególnych stanów, że s5, w którym znajduje się niezerowa nagroda, otoczony jest przez stany o malejących - wraz z odległością od s5 - wartościach V. Wskazany rozkład wartości jest ogólną cechą funkcji V. Cecha ta ujawnia się stopniowo wraz z kolejnymi interakcjami agenta z otoczeniem, który dysponując tego typu funkcją, może wyznaczyć optymalną strategię zachowania, czyli jest w stanie tak dobierać akcje, by zawsze pozyskać nagrody za pomocą minimalnej liczby kroków. W tym celu musi on „zawiesić" początkową strategię, za pomocą której eksplorował środowisko i zastąpić ją strategią zachłanną (*greedy*) opartą na funkcji wartości V (jest to tzw. tryb eksploatacji). Strategia zachłanna charakteryzuje się tym, że agent wybiera zawsze akcje, które powodują przejście do stanu o wyższej funkcji wartości, uwzględniając równocześnie rozkład nagród w eksplorowanym środowisku. Formalnie tego typu strategię dla środowiska w pełni deterministycznego definiuje się następująco:

$$\pi(s) = \arg\max_{a} [\, r(s,a) + \gamma\, V(s)],$$

gdzie V(s) jest funkcją wartości stanu, do którego agent przejdzie po wykonaniu akcji a. Uzyskujemy, odnosząc powyższą definicję do naszego przykładu, następujące możliwości:

$\pi(s1)$ = arg max [r(s2, a1) + γV(s2)]

= arg max [0 + 0,8*0,58] =**a1**

$\pi(s2)$ = arg max [r(s1, a1) + γV(s1); r(s2, a2) + γV(s1)]

= arg max [0 + 0,8*0,38; 0 + 0,8*0,78] = **a1**

π(s3) = arg max [r(s2, a1) + γV(s2); r(s4, a2) + γV(s4)]

= arg max [0 + 0,8*0,58; 0 + 0,8*0,99] = **a1**

π(s4) = arg max [r(s3, a1) + γV(s3); r(s5, a2) + γV(s5)] = **a1**

= arg max [0 + 0,8*0,78; 1 + 0,8*0,0] = **a1**

π(s7) = arg max [r(s6, a2) + γV(s6)]

= arg max [0 + 0,8*0,99] = **a2**

π(s6) = arg max [r(s7, a1) + γV(s7); r(s5, a2) + γV(s5)]

= arg max [0 + 0,8*0,4; 1 + 0,8*0] = **a2**

Powyższe zestawienie wyraźnie pokazuje, które akcje będą wybierane w poszczególnych miejscach korytarza. Co szczególnie interesujące, przedstawiony algorytm można w prosty sposób uzupełnić o elementy niedeterministyczne oraz niestacjonarne. Brak determinizmu znaczy, że ta sama akcja może spowodować, że agent z pewnym prawdopodobieństwem przejdzie do stanu x, a z pewnym do y. Z kolei brak stacjonarności znaczy, że informacja zwrotna w formie nagrody nie musi być taka sama. Zdarza się, że agent przy przechodzeniu z jednego stanu do innego czasami otrzyma nagrodę o wartości r1, a czasami o wartości r2. Agent posługujący się tego typu algorytmem, pomimo tych „utrudnień", nadal jest w stanie wyznaczyć optymalną strategię zachowań.

Z przedstawionych dotychczas analiz wynika, że błąd predykcji nagrody może po zrealizowaniu danego zachowania przyjąć trzy typy wartości:

- $\Delta>0$ – uzyskany stan oceniony został jako bardziej wartościowy, niż się spodziewano,
- $\Delta<0$ – uzyskany stan oceniony został jako mniej wartościowy, niż się spodziewano,
- $\Delta\approx0$ – uzyskany stan potwierdził wcześniejsze oczekiwanie.

Wszystkie powyższe typy udało się odnieść do zaobserwowanych w eksperymentach Schultza wyładowań neuronów dopaminergicznych. Odpowiednio: pierwszemu typowi

odpowiada wzrost wyładowań, drugiemu ich spadek, a trzeciemu tzw. poziom odniesienia (Colombo, 2014; Schultz i in., 1997). Algorytm TDRL, będący modelem obliczeniowym wymienionych typów fluktuacji, pozwala również zrozumieć, skąd bierze się efekt przeniesienia oraz blokowania w przypadku uczenia instrumentalnego. W sytuacji, gdy dany bodziec skojarzony zostanie ze stanem świata prowadzącym do pozyskania nagrody wówczas dojdzie do wyładowania neuronów dopaminergicznych, tak jak przewiduje to formuła obliczania błędu predykcji nagrody (Antonio Rangel i in., 2008). Funkcja wartości zostanie wówczas tak zaktualizowana, by w przyszłości, w podobnych okolicznościach, doprowadziła ona do wyboru akcji pozwalającej uzyskać nagrodę. Przedstawiona powyżej zasada działania algorytmu RL pozwala wyjaśnić, w jaki sposób obliczany jest błąd predykcji nagrody oraz – jakie elementy wpływają na jego wartość. Fakt, iż udało się powiązać wzorce wyładowań neuronów dopaminergicznych z parametrem Δ, spowodował, że przed neuronaukowcami pojawiła się niepowtarzalna okazja pozwalająca szerzej spojrzeć na kwestię procesów organizujących zachowania celowe. Skoro mamy podstawy sądzić, iż aktywność określonych struktur mózgowych opisać można za pomocą modelu obliczeniowego, zgodnego z algorytmem RL, to możemy również podjąć próbę zastosowania tego algorytmu do charakterystyki różnych typów naszych zachowań i oszacować, w przypadku których typów zachowań model ten pozwala na trafne ich wyjaśnienie lub przewidywanie.

Cechy algorytmu RL

Algorytm RL posiada trzy szczególne własności, które, jak się wydaje, można w naturalny sposób odnieść do zachowań ludzi i zwierząt. Są to odpowiednio: (1) definiowanie złożonych celów za pomocą nagród, (2) optymalizowanie strategii pozyskiwania nagród bez konieczności posiadania modelu środowiska oraz (3) brak funkcji planowania. Nietrudno zauważyć, że wymienione cechy w dużym stopniu charakteryzują dyspozycje dzieci we wczesnych fazach rozwoju ontogenetycznego.

> *Badania wykazały, że neurosensoryczne i neuromięśniowe struktury ciała [dzieci] rozwijają się przez spełnienie czynności, a podejmowane funkcje organizują się na coraz wyższym poziomie. Już od początku są jednak złożone i zintegrowane ze sobą (np. intersensoryczne powiązania).* (Harwas-Napierała & Trempała, 2004, s. 17).

Podobnie, jak się wydaje, przebiega proces nabywania nowych zdolności psychomotorycznych u noworodków. Na tym etapie rozwoju dziecko dysponuje już dobrze rozwiniętym centralnym układem nerwowym, który zawiera wszystkie kluczowe struktury mózgowe, w tym szlak mezolimbiczny i mezokortykalny, obsługujące mechanizm uczenia się ze wzmacnianiem. Dziecko wyposażone w tego typu struktury oraz odpowiednie systemy regulacji (odpowiedzialne m.in. za utrzymanie homeostazy) dysponuje mechanizmami nie tylko pozwalającymi zaspokajać podstawowe potrzeby, ale również uczyć się coraz bardziej złożonych zachowań celowych. Na tym etapie rozwoju opis poszczególnych pragnień i zamiarów dziecka w kategoriach pozyskiwanych nagród wydaje się być całkowicie naturalny. Zdobycie pokarmu, wody, potrzeba bliskości i bezpieczeństwa, zaspokojenie ciekawości – to przykłady naturalnych gratyfikacji, które mogą być włączone w model RL i w ten sposób mogą definiować cele dziecka, pojawiające się w toku rozwoju ontogenetycznego. Kiedy nagroda zostanie pozyskana, np. głód zostanie zaspokojony, wówczas prowadzące do tego celu zachowania oraz pozyskane informacje staną się częścią strategii, która z czasem zostanie zoptymalizowana i uogólniona. Na tym etapie rozwoju trudno jest mówić o planowaniu, a tym bardziej o deliberacji, która wymaga posiadania określonego modelu świata (patrz: punkt 2.2 w rozdziale 2.), dlatego, że tego typu możliwości pojawią się dopiero z czasem – w kolejnych fazach rozwoju, gdy dziecko będzie dysponowało złożoną siecią stanów intencjonalnych. Warto przypomnieć w tym miejscu, że warunkiem możliwości zaistnienia stanów intencjonalnych współtworzących wspomnianą sieć, zdaniem amerykańskiego filozofa, jest odpowiednio bogaty zbiór dyspozycji tła nabywanych przez dzieci m.in. metodą prób i błędów w związku z realizacją potrzeb poznawczych (Piaget, 1966; Searle, 1983) (z perspektywy uczenia się ze wzmacnianiem można tego typu potrzeby postrzegać jako specyficzny cel). W ten sposób nabywane są takie umiejętności jak: trzymanie przedmiotów, podpieranie się, wyciąganie dłoni, podciąganie się, pełzanie, raczkowanie, chodzenie, itd. (Harwas-Napierała & Trempała, 2004, s. 53–55). Dziecko potrafi spontanicznie włączać tego typu elementarne umiejętności w realizację celów biologicznych, które dotychczas były zaspokajane za pośrednictwem opiekunów. Widać zatem, że nabywane zdolności nie są „sztywno" związane z poszczególnymi celami, jak to ma miejsce w przypadku systemów sztucznych. Przeciwnie, w spontaniczny sposób są one łączone z wcześniejszymi doświadczeniami i zyskują coraz bardziej złożoną formę. Na obecnym etapie badań trudno jednoznacznie ustalić, które składowe systemu odpowiedzialnego za kontrolę działań intencjonalnych są obsługiwane przez mechanizm

uczenia się ze wzmacnianiem, a które wymagają dodatkowych dyspozycji, np. integracji z procesami poznawczymi (Shephard i in., 2014). Jedno wydaje się pewne: nabywanie poszczególnych umiejętności przez dzieci przebiega na poziomie behawioralnym niemal identycznie, jak w przypadku sztucznych systemów posługujących się algorytmem RL. Jeśli pod uwagę weźmiemy tylko umiejętności motoryczne dziecka, pominiemy procesy poznawcze odpowiedzialne za akwizycję języka oraz rozwój teorii umysłu, to okaże się, że kontrola zachowań w takim przypadku dokonuje się w głównej mierze poprzez mechanizm uczenia się ze wzmacnianiem (Zimbardo i in., 2010, s. 120). Potwierdzeniem tej tezy są osiągnięcia w dziedzinie systemów sztucznych, w szczególności w robotyce, która z powodzeniem stosuje algorytmy RL w tak złożonych zastosowaniach jak: umiejętność wykonywania akrobacji lotniczych (Ng i in., 2006) czy sterowanie samochodami autonomicznymi (Zhang i in., 2018). Wykorzystywanie takich samych lub bardzo podobnych mechanizmów uczenia się sekwencjonowania zachowań nie znaczy, że dziecko i systemy sztuczne wykształcą w podobnych warunkach identyczne strategie doboru działań.

Nie ulega wątpliwości, że różnice „w wyposażeniu" ludzi i robotów w systemy percepcyjne, motoryczne (na poziomie zachowań elementarnych) oraz ewaluacji nagród prowadzić będą do odmienności w zachowaniach tych pierwszych względem zachowań tych drugich. Fakt, iż obecne systemy sztuczne są na ogół jednocelowe, a ludzie i zwierzęta to agenci wielocelowi, również ma niebagatelny wpływ na postać przyjmowanych strategii. Nietrudno przewidzieć, jakie wyniki osiągnie agent, który będzie koncentrować się wyłącznie na pojedynczym celu, natomiast jest to bardzo trudne w przypadku agenta realizującego wiele różnorodnych celów. Efekt specjalizacji widać na przykład wśród wyczynowych sportowców, którzy przez setki powtórzeń uzyskują wyniki niedostępne dla zwykłych śmiertelników. Przedstawione różnice prowadzą do następującej konkluzji: trudno w pełni określić możliwości mechanizmu uczenia się ze wzmacnianiem na podstawie wyników eksperymentów opartych na prostych, kilkukrokowych sekwencjach zachowań (np. Pawłowa czy Skinnera). Pełen potencjał algorytmów RL ujawnia się najwyraźniej w działaniu systemów sztucznych, w przypadku których możemy bez „zniekształceń" obserwować, jak bogate i złożone mogą być sekwencje zachowań agenta, o ile pozwoli mu się na swobodną eksplorację środowiska i wielokrotne realizowanie tych samych celów. W takich warunkach metoda uczenia się ze wzmacnianiem w pełni ujawnia swój adaptacyjny i optymalizacyjny charakter.

Oprócz neurobiologicznych danych eksperymentalnych (Schultz i in., 1997) oraz potwierdzonego w badaniach behawioralnych podobieństwa pomiędzy procesem uczenia się dzieci i systemów sztucznych stosujących algorytm RL, warto również wskazać na argument ewolucyjny. Pojawienie się mechanizmu uczenia się ze wzmacnianiem jest, zdaniem Reada Montague, ewolucyjną „odpowiedzią" na niedającą się przewidzieć zmienność warunków środowiska, w którym żyją osobniki różnych gatunków. (Montague, 2006, s. 72). Przetrwanie, w sytuacji pojawienia się nowych, istotnych dla organizmu czynników w środowisku, wymaga redukcji niepewności poprzez odpowiednie przetwarzanie napływających informacji.

Algorytm RL jest wyjątkowo skutecznym narzędziem w realizacji tego typu zadania. Nie dziwi zatem fakt, że u wielu zwierząt wykształciła się zdolność do uczenia się na podstawie wzmocnień. Klasyczny eksperyment Pawłowa (Montague, 2006, s. 206), związany z odruchami warunkowymi jest przykładem na to, jak zwierzę potrafi wykorzystać w celu zdobycia nagrody dodatkową informację związaną z bodźcem warunkującym. Tego typu zjawisko – z perspektywy algorytmu RL – to nic innego, jak tzw. efekt blokowania polegający na odpowiednim dostosowaniu funkcji wartości, aby w przyszłości stan zapowiadający nagrodę wpływał na wybór zachowania prowadzącego do stanu z nagrodą. Podobne efekty można zaobserwować u małp, myszy, szczurów, delfinów i innych zwierząt (Montague, 2006, s. 98). Stwierdzić zatem można, że mamy solidnie uzasadnione przesłanki, by uznać mechanizm uczenia się ze wzmacnianiem za jedną z fundamentalnych form nabywania podstawowych zdolności motorycznych niezbędnych do realizacji zachowań celowych. Mechanizm ten jest wspólny dla obszernej klasy agentów biologicznych, natomiast osobniki gatunku *homo sapiens* posługują się nim także przy podejmowaniu działań intencjonalnych. W tym kontekście pojawia się ważne pytanie: czy wskazane cechy algorytmu RL pozwalają uznać go za wystarczający do wyjaśniania tego typu działań?

3.3 Algorytm RL jako podstawa złożonych działań intencjonalnych

W lutym 2015 roku w czasopiśmie *Nature* ukazał się artykuł zatytułowany *Human-level control through deep reinforcement learning* (Mnih i in., 2015), w którym zespół naukowców i programistów z firmy DeepMind opublikował, na podstawie analiz dotyczących efektywności zachowań różnego rodzaju postaci z „klasycznych" gier Atari,

wyniki badań odnoszących się do skuteczności algorytmu uczenia się ze wzmacnianiem. Wśród testowanych gier znalazły się następujące produkcje: „Breakaout", „PacMan", „River ride" czy „Moon patrol". Uzyskane wyniki potwierdziły, że odpowiednio wzbogacona wersja algorytmu RL (tzw. *deep reinforcement learning*) osiąga rezultaty porównywalne z najlepszymi graczami, a w niektórych przypadkach – nawet lepsze. Warto podkreślić, że skonstruowany system dysponował dokładnie takimi samymi danymi wejściowymi, jak człowiek – był to *de facto* zbiór pikseli pobierany z częstotliwością 50 Hz z ekranu monitora, na którym konsola Atari prezentowała obraz. Identyczny był również sposób posługiwania się joystickiem oraz dostęp do informacji o uzyskiwanych punktach. Z przedstawionych wyników zespołu kierowanego przez Demisa Hassabisa wynika, że tak dobre rezultaty okazały się możliwe za sprawą odpowiednio złożonej metody reprezentowania stanów świata oraz funkcji wartości w algorytmie RL. Do tego celu wykorzystano tzw. głęboką neuronową sieć splotową (*convolutional deep neural network*), która umożliwiła stworzenie funkcji mapującej zbiory pikseli – odnoszące się do zarejestrowanych kilkuklatkowych sekwencji ekranów z gry – na funkcję wartości *Q(s,a)* reprezentującą przewidywaną, zdyskontowaną sumę przyszłych nagród. Funkcja Q(s,a), pełni w algorytmie DRL analogiczną rolę, jak funkcja V(s) w opisanym powyżej algorytmie TDRL tzn. określa wartość spodziewanych w przyszłości nagród przy założeniu, że agent począwszy od stanu *s* będzie działał zgodnie ze strategią zachłanną, a jego pierwszym działaniem, które podejmie będzie działanie *a*. Agent, dysponując tego typu funkcją oraz stosując strategię zachłanną, potrafi optymalnie realizować zakładany cel (np. przejść do kolejnego etapu w grze PacMan). Przyjęte rozwiązanie, co szczególnie interesujące, uzyskało wysoki stopień ogólności. Ta sama konfiguracja systemu była stosowana do różnych gier, a mimo to – niemal w każdej z nich – system uzyskiwał poziom mistrzowski po odpowiednim (1-2 dniowym) treningu.

Ten spektakularny sukces wyraźnie pokazał, jak efektywnym narzędziem może być odpowiednio zaawansowana implementacja algorytmu RL. W 2018 roku podobnej klasy system po raz pierwszy w historii pokonał mistrza świata w GO („AlphaGo versus Lee Sedol", 2019), ostatnią grę planszową, w której człowiek skutecznie konkurował ze sztuczną inteligencją. Tego typu przykłady prowokują do zadania następującego pytania: skoro wiarygodny biologicznie algorytm RL, oparty na metodzie prób i błędów, jest w stanie uzyskać tak znakomite rezultaty w najróżniejszych dziedzinach, to czy możemy uznać, że jest on w stanie wyjaśnić wszystkie nasze zachowania, w tym działania

intencjonalne? Z teoretycznego punktu widzenia byłaby to pożądana sytuacja. Jeden mechanizm, dla którego dysponujemy precyzyjnym modelem obliczeniowym, wyjaśniałby całą różnorodność naszych zachowań. Niestety, choć osiągnięcia uczenia maszynowego są coraz bardziej spektakularne, to wciąż trudno uzyskać efektywność, którą obserwujemy w przypadku zachowań ludzkich, w szczególności tych, które realizowane są we współczesnym, zaawansowanym technicznie środowisku[39]. Gdyby nawet przyjąć, że RL pozwala na budowanie wielocelowych systemów zdolnych do radzenia sobie z różnego typu zagadnieniami i zadaniami, to nadal wiele spośród problemów skutecznie rozwiązywanych przez ludzi, będzie poza zasięgiem sztucznych systemów (Asokan, 2016). Można, jak się wydaje, nie wchodząc w szczegóły debaty dotyczącej ograniczeń współczesnych rozwiązań z obszaru sztucznej inteligencji, zidentyfikować elementy, które decydują o dodatkowej komplikacji ludzkich zachowań w odniesieniu do sztucznych systemów oraz do zwierząt. Odpowiednio złożona sieć stanów intencjonalnych, której istotną część możemy wyrazić w języku, a co za tym idzie – możemy w zaawansowany sposób „manipulować" jej składowymi, w znaczący sposób poszerza gamę rozwiązywalnych przez nas problemów, a w konsekwencji również poszerza repertuar dostępnych nam zachowań. Niestety, w „klasycznych"[40] implementacjach algorytmu RL nie podejmuje się prób wykorzystania wiedzy w formie symbolicznej, przy pomocy której, można by reprezentować wybrane fragmenty sieci stanów intencjonalnych. Należy zatem przyjąć, że na obecnym etapie badań metoda ta **nie jest** adekwatnym modelem złożonych działań intencjonalnych.

W dalszej części niniejszego rozdziału przedstawię najważniejsze założenia i cechy wybranych podejść, w których próbuje się rozszerzyć podstawową wersję metody uczenia się ze wzmacnianiem. Wśród tego typu propozycji na uwagę zasługuje hipoteza *superpower* Reada Montague, specjalizującego się w stosowaniu metod obliczeniowych do

[39] Rozróżnienie na tzw. wąską (*narrow*) i ogólną (*general*) sztuczną inteligencję dobrze obrazuje zasygnalizowane ograniczenie. O ile spektakularne osiągnięcia wąskiej sztucznej inteligencji (np. uczenia głębokiego) nie budzą wątpliwości (patrz: systemy rozpoznawania i segmentacji obrazów, przetwarzanie języka naturalnego, systemy kontroli agentów w świecie gier komputerowych, itp. (Desmond, 2019)), o tyle prace nad architekturą ogólnej sztucznej inteligencji nie mogą pochwalić się tego typu rezultatami. Wielość proponowanych rozwiązań, będących często uogólnieniem rozwiązań wąskich, wskazuje na wstępny charakter prowadzonych badań (Goertzel & Pennachin, 2007). W tym kontekście zdolność człowieka do funkcjonowania w bardzo złożonym środowisku kulturowo-technicznym, realizującego dziesiątki różnych celów, wydaje się być ciągle unikatowa i niedostępna sztucznym systemom. Nie znaczy to, że tego typu systemy w dobrze zdefiniowanych i kontrolowanych domenach nie będą w stanie działać równie skutecznie, co człowiek (patrz: debata dotycząca zagrożeń dla rynku pracy w związku z coraz większymi możliwościami automatyzacji ludzkiej pracy).

[40] Do klasycznych implementacji uczenia się ze wzmacnianiem zaliczam następujące algorytmy: TD-λ, Q-learning, SARSA.

badania mózgu. Z jej pomocą Montague wyjaśnia sposób, w jaki układ dopaminergiczny, implementujący metodę uczenia się ze wzmacnianiem, realizuje cele pozabiologiczne, które opierają się na przekonaniach. Informatycy, obok biologów i neuroobliczeniowców, również poszukują metod na włączenie wiedzy symbolicznej w algorytm RL. Istnieje szereg ciekawych rozszerzeń, które modyfikują proces uczenia się, uzupełniając go o dodatkową wiedzę wyrażoną w formie planów. Wreszcie, możliwe jest również potraktowanie sieci stanów intencjonalnych jako niezależnego podsystemu kontrolującego podsystem doboru zachowań oparty na algorytmie RL. Tego typu architektura, jak się wydaje, pozwala w naturalny sposób obsłużyć funkcję planowania symbolicznego, nieobecnego w podstawowych wersjach algorytmu RL, a zarazem wyjaśnić, jak dochodzi do procesu stopniowego automatyzowania zachowań. Wskazane rozszerzenia, a przynajmniej pewne ich elementy, posłużą w ostatnim rozdziale do skonstruowania zintegrowanego modelu działań intencjonalnych.

Hipoteza „nad-mocy”[41]

Nie mamy wątpliwości, obserwując, jak drapieżnik powoli zbliża się pośród traw sawanny do swej potencjalnej ofiary, że zwierzę realizuje pewien jasno określony cel – że chce zdobyć pożywienie. Podobną jednoznacznością charakteryzują się, zdaniem Montague’a, jeszcze cele związane z prokreacją, bezpieczeństwem czy – szerzej – z homeostazą. Ich uniwersalność w świecie przyrody wynika wprost z zasad ewolucji. Większość obserwowanych zachowań zwierząt ma z nimi bezpośredni lub pośredni związek. Podobnie jest również w przypadku gatunku ludzkiego, wiemy jednak doskonale, że człowiek potrafi zawiesić tego typu podstawowe potrzeby i świadomie działać wbrew instynktowi samozachowawczemu. Kamikadze czy fundamentaliści muzułmańscy – to przykłady sprawców ataków samobójczych, którzy w imię określonych przekonań gotowi byli zginąć, by osiągnąć założone na ich podstawie cele (Lasota & Grenda, 2017). W tym kontekście powstaje zatem pytanie: czy dysponujemy takim rozumieniem mechanizmów osiągania celów, które pozwalałoby wyjaśnić zarówno zachowania związane z pozyskiwaniem pożywienia, jak i np. ze świadomym prowadzeniem głodówki z przyczyn politycznych czy samobójczym atakiem terrorystycznym?

Od razu na wstępie należy rozróżnić dwa typy strategii definiowania celów: (1) jawną — preskryptywną oraz (2) pośrednią – za pomocą sygnałów nagrody (tzw. hipoteza

[41] Obszerne fragmenty niniejszego rozdziału zostały opublikowane w artykule: (M. Cichosz, 2010).

nagrody). Pierwsza odmiana strategii polega na zaplanowaniu wszystkich zadań niezbędnych do osiągnięcia stanu pożądanego z perspektywy realizowanego celu. Mamy w tym przypadku do czynienia z przepisem, który określa kroki wymagane do tego, aby cel został osiągnięty. Z preskrypcją spotykamy się na przykład podczas instalacji programu komputerowego. Osiągamy pożądany stan, czyli oprogramowanie zainstalowane na komputerze, wykonując następujące działania: (1) akceptacja instalacji i regulaminu, (2) wskazanie docelowego katalogu, (3) uruchomienie przegrywania plików, itd. Problem polega jednak na tym, że tego typu opis celu zakłada posiadanie przez użytkownika precyzyjnego modelu środowiska, w którym będzie on realizowany, modelu reprezentującego różne sytuacje, wyjątki, błędy, czyli tzw. dynamikę. Podejście preskryptywne wymaga zatem niezwykle dokładnej znajomości dziedziny. Wszędzie tam, gdzie mamy do czynienia z dużą zmiennością, np. w świecie przyrodniczym, taka reprezentacja celu jest w zasadzie bezużyteczna. Jeśli w ogóle występuje, to jest raczej wyjątkiem niż regułą.

Druga forma definiowania celu opiera się na wspomnianej wcześniej hipotezie nagrody. Zgodnie z tą hipotezą dowolny cel można wyrazić w formie strategii wyboru zachowań, która maksymalizuje zdyskontowaną sumę przyszłych nagród. Nagroda w tym kontekście definiowana jest przez pozytywną lub negatywną wartość, którą agent przypisuje pewnemu obiektowi (np. pokarmowi), pewnemu aktowi behawioralnemu (np. rytualnym zachowaniom godowym) lub wewnętrznemu stanowi organizmu (np. stresowi) (Schultz i in., 1997). Natomiast z perspektywy informacyjnej nagrodę można traktować jako tzw. przekaz zwrotny wygenerowany przez środowisko na skutek zrealizowanego zachowania. Wiemy, na podstawie badań nad zwierzętami, że ten sposób definiowania celów jest powszechny w świecie przyrody (Montague, 2006, s. 55; Schultz i in., 1993; Wightman, 2006). Nie jest to oczywiście przypadek. Wiemy dokładnie, na podstawie analiz modeli uczenia się ze wzmacnianiem, dlaczego tak się dzieje. Algorytm uczenia się ze wzmacnianiem wypracowuje na podstawie rzeczywistych lub hipotetycznych doświadczeń, krok po kroku, optymalną strategię doboru zachowań zapewniającą organizmowi długoterminowe korzyści, czyli zdobywanie nagród.

Montague traktuje powyższe rozróżnienie jako ważny argument w dyskusji na temat mechanizmu decydującego o kształcie naszych zachowań. Skoro cele oparte na hipotezie nagrody są bardziej efektywne, niż planowanie preskrypcyjne, to pojawia się pytanie, jak można wzbogacić mechanizm uczenia się ze wzmacnianiem, by objął on również

przypadki, w których kluczowe są czynniki kulturowe, będące często w konflikcie z czynnikami biologicznymi? Zjawisko to od wieków intrygowało twórców literatury, filozofów czy teologów (Hobbes, 2009; Rousseau, 1930; Wilson, 1988). Wciąż powraca pytanie o to, czy gotowość do oddania życia w imię określonych przekonań nie jest sprzeczna z biologiczno-ewolucyjnym „imperatywem" przetrwania? Wiemy, że wśród zwierząt dochodzi czasami do skrajnych form samopoświęcenia, ale mechanizm, który o tym decyduje, jest w pełni spójny ze strategią przetrwania oraz wymaganiem transmisji genów — przykładem mogą być mrówki, które w sytuacji zagrożenia potrafią eksplodować, by ochronić innych członków mrowiska (Oster, 1978, s. 226). Jednak w przypadku ludzi tego typu eksplikacja wydaje się mało przekonująca, trudno np. dopatrzeć się jakichś korzyści gatunkowych w zbiorowym akcie samobójstwa popełnionego przez członków sekty religijnej. Zdaniem Montague tego typu przypadki wskazują na następującą możliwość: odpowiednio rozwinięty układ nerwowy posiada zdolność nadawania szczególnej wartości określonym stanom intencjonalnym – uzyskują one status bardzo cennych nagród. Znaczy to, że przestrzeń potencjalnych celów jest otwarta i nie ogranicza się tylko do obiektów, zachowań i wyznaczonych przez uposażenie genetyczne stanów wewnętrznych organizmu. Co szczególnie interesujące, tego typu status mogą uzyskać tak abstrakcyjne idee, jak „przejście na wyższy poziom świadomości za pośrednictwem komety Hale'a-Boppa" czy „strajk głodowy w imię wolności politycznej" (Montague, 2006, s. 110). Nie są jeszcze przebadane zasady działania mechanizmu odpowiadającego za nadawanie abstrakcyjnym ideom statusu nagrody. Wiemy tylko, że zwierzęta do pewnego stopnia również potrafią wyhamowywać impulsywne reakcje powiązane z takimi potrzebami jak głód czy pragnienie. Zachowania rekinów są w większym stopniu uwarunkowane genetycznie, niż psów, w przypadku których można zidentyfikować zachowania świadczące o ograniczonej, ale jednak obserwowalnej zdolności hamowania celów czysto biologicznych, np. zdolności do czasowego powstrzymywania się od jedzenia w przypadku, gdy dostarczona karma nie spełnia oczekiwań „rozpieszczonego pupila".

Jakie są konsekwencje faktu, iż abstrakcyjna idea może stać się nagrodą? Znaczy to, że tego typu reprezentacja zacznie wpływać na sposób przetwarzania informacji pozyskiwanych ze środowiska. W przypadku każdego celu realizowanego za pomocą mechanizmu uczenia się ze wzmacnianiem będzie dochodziło w trakcie wykonywanych zadań do oceny napływających danych: na ile prowadzą one do korzystnych,

niekorzystnych lub neutralnych stanów świata oraz – czy są zgodne z długoterminowymi przewidywaniami. Nie wiemy obecnie, w jaki sposób przebiega tego typu ocena. Potrafimy jednak na podstawie fluktuacji neuronów dopaminergicznych zarejestrować fakt, że zachodzi tego typu ewaluacja. Świadczą o tym eksperymenty dotyczące tak „niebiologicznych" zachowań, jak inwestowanie na giełdzie (Knutson i in., 2001; Montague, 2006, s. 178) czy wybór produktu na podstawie marki (McClure i in., 2004). Podobne obserwacje poczyniono w odniesieniu do abstrakcyjnych uczuć społecznych, jak np. zaufanie do innej osoby lub poczucie straty związane z błędną decyzją (Braver & Cohen, 2000). Innymi słowy, kiedy bada się tego typu społeczne, zdeterminowane kulturowo zachowania, to obserwujemy taką samą aktywność neuronów dopaminergicznych, jak w przypadku, gdy towarzyszą one osiąganiu typowych celów biologicznych, np. zdobywaniu pożywienia w związku z odczuwaniem głodu. Wydaje się jednak, że nawet, gdy abstrakcyjna idea uzyska status nagrody i od tej pory system realizacji celów co jakiś czas będzie dążył do jej pozyskania, to i tak trudno zrozumieć, dlaczego tego typu nagroda miałaby prowadzić do tak niebezpiecznych dla organizmu zachowań, jak długotrwała głodówka wyniszczająca organizm. Odpowiedź tkwi, zdaniem Montague, w działaniu dwóch mechanizmów: (1) mechanizmu nadawania celowi wartości priorytetowej oraz (2) mechanizmu ustanawiania i podtrzymywania celów. Pierwszy z nich odpowiada za zapewnienie dostępności celów biologicznych. Związane z nimi nagrody mają bowiem szczególne znaczenie dla dobrostanu organizmu i dlatego interpretowane są jako niezwykle wartościowe. W konsekwencji, gwarantuje to, że ich pozyskiwanie będzie zawsze realizowane efektywnie (niemal optymalnie), a w sytuacji konfliktu typu: ukraść coś i zaspokoić silny głód czy być uczciwym, wybrany zostanie prawdopodobnie ten pierwszy wariant. Cele podstawowe, oprócz wysokiej wartości nagród (*primary rewards*), jak określa je Montague, wyróżnia jeszcze jedna ważna właściwość, mianowicie – cykliczność. Co jakiś czas każda z wymienionych potrzeb „instalowana" jest w systemie osiągania celów i wszystkie działania „orientowane" są na jej zaspokojenie. Za „zainstalowanie" nowego celu odpowiada mechanizm ustanawiania i utrzymywania celów (*goal setting*), który w trybie ciągłym monitoruje napływające informacje z otoczenia i „decyduje" o ewentualnej zmianie lub kontynuacji aktualnie realizowanego celu (Montague, 2006, s. 128). Analizowane informacje mogą mieć zarówno źródło zewnętrze (środowisko), jak i wewnętrzne (stany energetyczne organizmu lub stany mentalne). W każdym z wymienionych przypadków podlegają one ocenie przez ten sam podsystem dopaminowy, który wykorzystywany jest przez mechanizm realizowania celów.

Podsumowując, cele podstawowe cechują się wysoką wartością nagród oraz cyklicznością. Zdaniem Montague również cel związany z realizacją określonej idei, np. z pragnieniem zdobycia Mont Everestu może uzyskać status podobny do podstawowego celu biologicznego. W takim przypadku wiele zachowań będzie organizowanych w taki sposób, aby pomagały w realizowaniu tego typu pragnienia. Z czasem, w wyniku wzmocnień może ono uzyskać tak wysoką wartość, że będzie w stanie „zawetować" dowolny inny cel, w tym również biologiczny cel podstawowy. Tego typu sytuacja może w określonych okolicznościach doprowadzić nawet do samozagłady. Uzyskanie przez ideę (abstrakcyjną reprezentację) statusu nagrody podstawowej jest czymś wyjątkowym i unikatowym dla gatunku ludzkiego. Montague proponuje nazwać tego typu dyspozycję mianem „nad-mocy" (*superpower*) (Montague, 2006, s. 88). Ten amerykański neuronaukowiec twierdzi, że istnieje wiele zabezpieczeń, aby przypadkowa reprezentacja nie uzyskała tak szczególnej pozycji w hierarchii celów. Mogłoby to bowiem doprowadzić do wielu nieefektywnych zachowań, a w konsekwencji nawet do zagłady całego gatunku (Montague, 2006, s. 139). Zdolność do nadawania wysokiej wartości określonym ideom ma również pozytywny aspekt: pozwala skupić się na realizacji celów poznawczych, często niezależnych od potrzeb biologicznych. Nasze zachowania cechuje daleko posunięta elastyczność, innowacyjność oraz adaptacyjność. Wiąże się z tym ryzyko „przewartościowania" danej idei, włącznie z możliwością samounicestwienia.

Sformułowana w ramach hipotezy nad-mocy propozycja Montague'a, aby traktować określoną ideę (pewien stan intencjonalny) jako specyficzny rodzaj nagrody – jest oryginalna i intrygująca. Wynika z niej, że u podłoża tak ekstremalnych zachowań, jak poświęcenie własnego życia ze względu na przekonania polityczne, religijne, etyczne, itp. (Montague, 2006, s. 88) leży ten sam mechanizm doboru zachowań, który wspiera nas w zaspokajaniu podstawowych potrzeb biologicznych. Dostępny obecnie materiał empiryczny jest zbyt skromny, aby można było wiarygodnie orzekać o trafności tej hipotezy. Potwierdzają ją, zdaniem Montague, wzorce zachowań osób uzależnionych, dobrze obrazujące do jak ekstremalnych działań gotowi są ludzie, u których w wyniku przyjęcia narkotyku doszło do „rozregulowania" systemu pozyskiwania nagród. Badania na zwierzętach związane z podawaniem im heroiny wykazały, że w istotny sposób modyfikuje ona działanie neuronów dopaminergicznych (Olds, 1958). Za każdym razem, kiedy przyjmowana jest uzależniająca substancja, pojawia się wyrzut dopaminy w układzie nagrody. Poszczególne struktury mózgowe otrzymują, oprócz określonego zespołu doznań,

również informację o tym, że wystąpił błąd predykcji nagrody o wartości $\Delta>0$. Z perspektywy agenta jako systemu uczącego się implikuje to odbiór następującego komunikatu: „pozyskana nagroda jest większa niż oczekiwano". Ponieważ taki komunikat pojawia się za każdym razem, kiedy przyjmowany jest narkotyk, dlatego z perspektywy organizmu wartość tego typu substancji jest w pewnym sensie nieskończona. W konsekwencji, uzależnione zwierzę lub uzależniona osoba zaczynają całe swoje zachowanie podporządkowywać pozyskiwaniu narkotyku, nawet w sytuacji, gdy nie występują już wywoływane przez niego wcześniej stany euforii. Efekt blokowania wbudowany w algorytm RL powoduje, że stany świata (miejsca, ludzie, zdarzenia) – skojarzone z zażywaniem narkotyku – zaczynają funkcjonować analogicznie do wyzwalaczy (*triggers*) wywołujących nawroty choroby. Dlatego tak ważne jest, by uzależnione osoby całkowicie zmieniły środowisko i otoczenie, a nawet zerwały dotychczasowe kontakty. Widać zatem, że podobnie jak w przypadku przewartościowanych idei, również w odniesieniu do uzależnień obserwujemy podobny wzorzec zachowań – chęć pozyskania nagrody, często za wszelką cenę. W wielu przypadkach tego typu wzorzec jest dla organizmu całkowicie destrukcyjny, ale w pełni daje się opisać w perspektywie mechanizmu uczenia się ze wzmacnianiem.

Montague zdaje sobie sprawę, że powyższa koncepcja nie wyczerpuje złożoności zjawiska:

> *W rzeczywistości rzeczy nie są tak proste, jak to przedstawiono, problemy są często bardziej abstrakcyjne, a proste sygnały wzmocnień typu ciepło/zimno żałośnie niewystarczające. Ale mózg ma jeszcze kilka sztuczek w zanadrzu. Sygnały wzmocnień są bardziej złożone, istnieje więcej, niż jeden system oceny napływających informacji, który ewaluuje różne aspekty każdego działania, a nawet aktywuje wiele konkurencyjnych celów. Ta zdolność do porównywania celów, nadawania im odpowiedniej rangi i utrzymywania wielu aktywnych celów jednocześnie zależy w dużej mierze od naszej kory przedczołowej.*[42] (Montague, 2006, s. 110).

[42] „In the real world, things are not this neatly presented, the problems are often more abstract, and simple warmer-colder critic signals would be woefully insufficient. But the brain has a few more tricks up its sleeve. The brain's critic signals are more complex, there is more than one critic ranking different aspects of each action and even keeping multiple competing goals active. This ability to compare goals, rank goals, and keep multiple goals in mind depends in large part on our expanded prefrontal cortex. We will focus on critic signals, but the functions of the prefrontal cortex are an important subject because it connects some of the most interesting aspects of our mental lives to neural function" (Montague, 2006, s. 110).

Warto zauważyć, nie kwestionując wskazanej przez Montague'a prawidłowości, że trudno, ograniczając się jedynie do specyficznego rodzaju nagród, wyjaśnić ogromną różnorodność typów działań charakterystycznych dla człowieka funkcjonującego w złożonym środowisku cywilizacyjnym. Nawet jeśli uwzględni się występowanie hierarchii celów, wykorzystując zasady działania algorytmu RL, a także tam, gdzie to możliwe – uwzględni się równoległe sposoby ich realizacji, to i tak nie da się wyjaśnić dwóch kluczowych kwestii:

1. Jakiego rodzaju relacja zachodzi między ideą abstrakcyjną, posiadającą status nagrody, a siecią stanów intencjonalnych, która stanowi jej reprezentacyjną bazę?
2. W jaki sposób realizowane są cele, wyznaczane na podstawie wcześniej opracowanych planów, które są niezbędne do skutecznego funkcjonowania w złożonym otoczeniu (np. w środowisku miejskim albo w zinstytucjonalizowanej, zhierarchizowanej grupie społecznej)?

Idea abstrakcyjna jako nagroda a sieć stanów intencjonalnych

Uznając opracowaną przez Montague hipotezę nad-mocy za prawdopodobną, w naturalny sposób nasuwa się sformułowane powyżej pytanie: jaka jest relacja między ideami abstrakcyjnymi-nagrodami a siecią stanów intencjonalnych postulowaną przez Johna Searle'a? Niestety, Montague jest w tej kwestii nieprecyzyjny. Posługuje się on enigmatycznym pojęciem „idei abstrakcyjnej", którego nie wyjaśnia, nie precyzuje też ani typu, ani struktury tej formy reprezentacji. Trudno zatem orzec, czy status nagrody mogą uzyskać jedynie stany intencjonalne o nakierowaniu na zgodność typu świat→umysł (np. pragnienia, prior intencje, nadzieje, itp.), czy również stany o nakierowaniu na zgodność umysł→świat (np. przekonania)? Brak tego typu deklaracji wskazuje na nazbyt uproszczony charakter analizy Montague. Bez koncepcji dotyczącej funkcjonowania stanów intencjonalnych, łatwo można by uznać, że tego typu stany są jak niezależne, dobrze wyodrębnione, dyskretne jednostki. Przedstawione przez Searle'a argumenty przemawiają przeciwko tego typu ujęciu. Pragnienie wystartowania w wyborach prezydenckich wymaga od podmiotu, jak zauważa amerykański filozof, by ten dysponował całą siecią dodatkowych, komplementarnych przekonań, np. „Stany Zjednoczone to kraj, w którym co jakiś czas odbywają się wybory prezydenckie.", „By móc zostać prezydentem USA, potrzebna jest nominacja jednej z dwóch partii.", „Wymagane jest, by kandydat na prezydenta urodził się na terenie jednego ze stanów.", „Obywatel USA nie może

kandydować na prezydenta, jeśli wcześniej pełnił tego typu funkcję dwukrotnie.", itd. Przytoczony przykład dość jednoznacznie pokazuje, w jak złożonych kontekstach mogą funkcjonować idee abstrakcyjne Montague. Jeśli tego typu idea może – zgodnie z hipotezą nad-mocy – uzyskać status nagrody, to może to nastąpić jedynie za pośrednictwem szerszego intencjonalnego kontekstu, a nie w izolacji. Powyższe spostrzeżenie nasuwa kolejne pytanie: jaką funkcję w układzie odpowiedzialnym za pozyskiwanie nagród mogą pełnić stany należące do kontekstu – czy coś do niego wnoszą, czy są „niewidoczne" dla mechanizmu uczenia się ze wzmacnianiem?

Przypomnę, że w podstawowej wersji algorytm RL operuje jedynie pięcioma bazowymi typami reprezentacji: zachowaniami (a), stanami świata (s), nagrodami (r), funkcją wartości (V) oraz błędem predykcji nagrody (δ). Wydaje się, że w tak konceptualnie prostym układzie nie da się wyrazić złożonych zależności, które występują w sieci lub podsieci stanów intencjonalnych. Trudno w związku z tym zaakceptować podejście, które zaniedbuje wpływ tego typu kontekstu. Wydaje się zatem, że bez odpowiedniego rozszerzenia algorytmu nie da się włączyć w jego działanie informacji zgromadzonej w stanach intencjonalnych „współkonstytuujących" daną ideę. Należałoby się spodziewać, pamiętając o kosztach, jakie ponosi organizm w związku z tworzeniem, utrzymywaniem, reorganizowaniem oraz świadomym dostępem do tego typu stanów, że ewolucja „zadbała" również o to, by zgromadzona w ten sposób informacja była wykorzystywana w systemie kontroli zachowań. W oryginalnej propozycji Montague'a nie ma na ten temat żadnych wskazówek, jednak niedawno zespół z kierowanego przez niego laboratorium przeprowadził ciekawy eksperyment, który, jak się wydaje, potwierdza występowanie związku pomiędzy określonymi stanami intencjonalnymi a przebiegiem procesu decyzyjnego, wykorzystującego mechanizm uczenia się ze wzmacnianiem (Gu, X. i in., 2015). W artykule zatytułowanym *Belief about nicotine selectively modulates value and reward prediction error signals in smokers* przedstawiono wyniki eksperymentu z udziałem nałogowych palaczy. Zadanie polegało na tym, by zaproszeni ochotnicy – w trakcie sesji funkcjonalnego obrazowania metodą rezonansu magnetycznego (fMRI) – inwestowali niewielkie kwoty pieniędzy w wybrane spółki giełdy nowojorskiej, których notowania widzieli na ekranie (jest to tzw. *sequential choice task paradigm*). Każdy z uczestników eksperymentu miał przed sesją wypalić również specjalnego papierosa, po której to czynności informowano badanych o jego składzie. Jednym mówiono, że papieros zawierał nikotynę, a innym, że nie zawierał. Celem badania była ocena wpływu przekonań na przebieg podejmowanych decyzji. Podawana informacja była świadomie

zmanipulowana, tzn. czasami, gdy papieros zawierał nikotynę, informowano badanego, że był on bez nikotyny, a czasami, gdy jej nie zawierał, informowano, że papieros był z nikotyną (tzw. efekt placebo). Po przeanalizowaniu danych okazało się, że:

- przekonanie o wypaleniu papierosa z nikotyną wpływa na neuronalną odpowiedź struktury związanej z sygnałami nagrody r w taki sam sposób, jak faktyczne wypalenie papierosa z nikotyną; w konsekwencji, obserwowano również odpowiednie zmiany w zachowaniu, tzn. w sposobie inwestowania w akcje spółek,
- przekonanie o wypaleniu papierosa z nikotyną wpływa na neuronalną odpowiedź struktury związanej z błędem predykcji nagrody (Δ); również w tym przypadku obserwowano odpowiednie zmiany w zachowaniu.

Powyższe rezultaty wyraźnie pokazują, że określony stan intencjonalny, nawet jeśli nie jest związany bezpośrednio z procesem decyzyjnym, to modyfikuje sposób funkcjonowania układu odpowiedzialnego za uczenie się ze wzmacnianiem. Jest to przykład oddziaływania elementów sieci stanów intencjonalnych na mechanizmy powiązane z doborem zachowań. Można zatem z dużym prawdopodobieństwem stwierdzić, że analogiczne, a być może nawet silniejsze efekty powinniśmy obserwować w przypadku przekonań i innych stanów intencjonalnych współkonstytuujących treść danej idei abstrakcyjnej (Warto w tym miejscu zauważyć, że tego typu możliwość – z perspektywy psychologii społecznej, socjologii czy badań antropologicznych – może wydawać się czymś oczywistym, gdyż wszystkie wymienione dyscypliny próbują wyjaśnić związki między siecią przekonań a ludzkimi zachowaniami. Jednakże pojawiające się w tych sytuacjach poczucie oczywistości jest efektem „gruboziarnistości" wyjaśnień, a nie ich niedającej się zakwestionować trafności.). Tego typu spostrzeżenie daje asumpt do sformułowania następującego pytania: czy w ogóle możliwe jest włączenie do podsystemu działającego na podstawie algorytmu RL dodatkowej informacji zapamiętanej w formie stanów intencjonalnych, bez „zniekształcania" jego kluczowej cechy, mianowicie zdolności do znajdowania optymalnej strategii zachowań dla zadanego celu? Odpowiedź na powyższe pytanie przedstawię w dalszej części niniejszego rozdziału, przede wszystkim w paragrafie zatytułowanym *Wybrane rozszerzenia metody uczenia się ze wzmacnianiem*.

Inicjowanie celów abstrakcyjnych

Dla zachowania jasności wywodu w powyższych rozważaniach pominięto kwestię aktywowania wybranego celu. Taka uproszczona sytuacja jest typowa w przypadku działania systemów sztucznych. Dany system realizuje na ogół jeden dobrze zdefiniowany cel, który wyrazić można w postaci następującego stwierdzenia: pozyskać jak największą liczbę nagród określonego typu w przewidzianym przez projektantów środowisku, złożonym z szeregu dyskretnych stanów. Za pomocą tych dwóch składników – nagród i stanów środowiska, a także przy wykorzystaniu mechanizmu uczenia się – możliwe staje się stopniowe wyznaczenie optymalnej funkcji wartości, na podstawie której można określić optymalną strategię wyboru zachowań. Sytuacja znacząco się komplikuje w przypadku organizmów biologicznych, które w środowisku przyrodniczym muszą realizować co najmniej kilkanaście różnych celów, w tym przede wszystkim te, które umożliwiają przetrwanie: zdobycie pożywienia, znalezienie partnera, monitorowanie otoczenia w celu uniknięcia zagrożeń. Specyficzne dla tych podstawowych potrzeb typy nagród są cyklicznie aktywowane przez odpowiednie mechanizmy regulacyjne. Jest to zatem rodzaj automatyzmu, który na podstawie stanu organizmu załącza odpowiednią konfigurację RL zapewniającą pozyskanie nagrody i przywrócenie stanu równowagi.

Sytuacja się komplikuje, kiedy weźmie się pod uwagę determinanty pozabiologiczne. W tego typu przypadkach czynnikiem inicjującym określone zachowania są albo informacje pozyskane ze środowiska, albo stany wewnętrzne organizmu (np. przemożna chęć zrelaksowania się wywołana długotrwałym wysiłkiem umysłowym, np. uczeniem się do egzaminu). Istotne jest również to, by systematycznie monitorowane były napływające informacje. Łatwo sobie wyobrazić, co się stanie, kiedy idąc do pracy, zauważymy przy drodze jakiś cenny przedmiot. Zatrzymamy się i prawdopodobnie schylimy się po niego, aby przyjrzeć się mu dokładniej. W ten sposób aktualnie realizowany cel zostanie przerwany, a na jego miejsce załączony zostanie cel związany z dostrzeżoną rzeczą. Z drugiej strony, nie możemy – chcąc dotrzeć do pracy na wyznaczoną godzinę – ciągle zmieniać celów. Widać zatem, że mechanizm zarządzania celami jest podatny na modyfikacje, w których uwzględnia się nowe, napływające informacje. Równocześnie, mechanizm zarządzania celami musi być na tyle stabilny, żeby umożliwić organizmowi skuteczne realizowanie poszczególnych zamierzeń. Read Montague zaproponował koncepcję tego typu mechanizmu we wspomnianej wcześniej pracy zatytułowanej *Why choose this book?*. Przedstawiona przez niego koncepcja stanowi w dużym stopniu twórcze rozwinięcie hipotezy „bramkowania dopaminowego" (*dopamine gating hypothesis*), którą sformułowali trzej amerykańscy psychologowie: O'Reilly, Braver i Cohen (OBC)

(O'Reilly i in., 1999). Wymienieni badacze, korzystając głównie z symulacji obliczeniowych oraz dostępnych danych neurobiologicznych, doszli do wniosku, że podczas realizacji określonych zadań wymagających kontroli poznawczej (np. zadania AX-CT) istotna jest współpraca dwóch systemów mózgowych: kory przedczołowej (*prefrontal cortex* – PFC) oraz mechanizmu uczącego się, który nadzorowany jest przez struktury odpowiedzialne za produkcję dopaminy (DA). Uczeni ci wykazali, wprowadzając symulację, że DA jest w stanie stabilizować i stopniowo wzmacniać reprezentacje w PFC, które są aktywowane w związku z realizacją danego zadania/celu, czyli tzw. informacje związane z kontekstem (*contextual information*). Tego typu reprezentacje wpływają zwrotnie na efektywne pozyskiwanie nagród, a co za tym idzie – na realizację celu (por. Braver & Cohen, 2000). Z czasem badacze ci powiększyli wskazany układ PFC-DA o system pamięci długotrwałej (*hippocampus* – HCMP) oraz o odpowiednie struktury motoryczno-sensoryczne (*posterior perceptual and motor cortex* – PMC), tworząc całościową koncepcję pamięci roboczej rozumianej jako proces kontroli i koordynacji systemów mózgowych, niezbędnych do rozwiązywania zadań poznawczo-motorycznych (O'Reilly, Braver, & Cohen, 1999). Aby zrozumieć związki pomiędzy poszczególnymi systemami, należy rozważyć następujący przykład: wyobraźmy sobie, że chcemy znaleźć w korespondencji znajdującej się w naszej skrzynce mailowej imię dziecka przyjaciela ze studiów. Rok wcześniej przesłał je nam nasz wspólny znajomy. List znajduje się w skrzynce elektronicznej, którą podzieliliśmy na wiele folderów. Niestety, nie pamiętamy ani tytułu maila, ani nazwy folderu, w którym go umieściliśmy. W zaistniałej sytuacji postanawiamy przypomnieć sobie coś, co pomoże nam odnaleźć poszukiwane imię. Na przykład, przeglądamy wszystkie wiadomości od autora listu, który przesłał nam imię dziecka (jest to dobry kolega, znający przyjaciela ze studiów) oraz staramy się w przybliżeniu zawęzić okres przeszukiwań, przypominając sobie dodatkowe okoliczności, z którymi można by skojarzyć moment odebrania wiadomości (np. powrót z konferencji w Paryżu).

Zdaniem O'Reilly'a, Bravera i Cohena (1999) – tego typu informacje przechowywane są w hipokampie (HCMP), w którym następuje wiązanie określonych cech wspomnień i zapisanie ich w formie unikatowego epizodu. W powyższym przykładzie, informacja o imieniu dziecka powiązana została z otrzymaniem maila oraz z uczestnictwem w konferencji. Aby dotrzeć do tej wiadomości przystępujemy do przeglądania wybranych folderów oraz czytania znalezionych w nich maili, by za pomocą zawartych w nich

informacji – jak najsprawniej odszukać interesujące nas imię. Zdaniem wyżej wymienionych uczonych, tego typu przeszukiwanie możliwe jest dlatego, że reprezentacje pobrane z HCMP (autor maila i data konferencji) utrzymywane są w korze przedczołowej (PFC) jako aktywny stan – aż do zakończenia zadania. W tym czasie PFC „dba" o to, by informacje pozyskiwane w trakcie przeszukiwania maili nie wyparły tych, które w istotny sposób zawężają zakres przeszukiwania (kierują jego przebiegiem). Z kolei, za zdolności takie jak: czytanie informacji, obsługa myszki czy użytkowanie systemu pocztowego odpowiada system PMC, nadzorowany w dużym stopniu przez aktywowane w danym momencie reprezentacje PFC. Znaczy to, że stan pamięci wymaga ciągłej aktualizacji od momentu zainicjowania przeszukiwania. Przykładowo, kiedy wyświetli nam się pełna lista folderów, na ogół wybierzemy z niej mały podzbiór składający się z kilku elementów, na których skupimy naszą uwagę. Tak wyselekcjonowany podzbiór stanie się składową pamięci, aktywnie „organizującą" proces przeszukiwania. Tego typu mechanizm wymaga interakcji między reprezentacjami aktywowanymi na poziomie PFC, trwałą – zawartą w PMC – wiedzą dotyczącą funkcji poszczególnych folderów (skrzynka odbiorcza, skrzynka nadawcza, spam, itp.) oraz danymi przechowywanymi w HCMP, odnoszącymi się do okoliczności umieszczenia maili w folderach oraz ich treści. Stopniowo pamięć robocza obejmie wszystkie niezbędne informacje po to, by odnaleźć poszukiwany mail. Znaczy to, że w zależności od potrzeb danego zadania aktywowane i dezaktywowane są elementy w pamięci nadzorowanej przez PFC. Proces szukania imienia dziecka zakończy się, gdy przeglądając poszczególne foldery oraz treści maili, porównując daty i listy autorów z tymi, które są w aktywnej pamięci, zostanie odnaleziona odpowiednia informacja (por. O'Reilly i in., 1999). Następujący diagram obrazuje wymienione w przykładzie systemy oraz istniejące między nimi relacje:

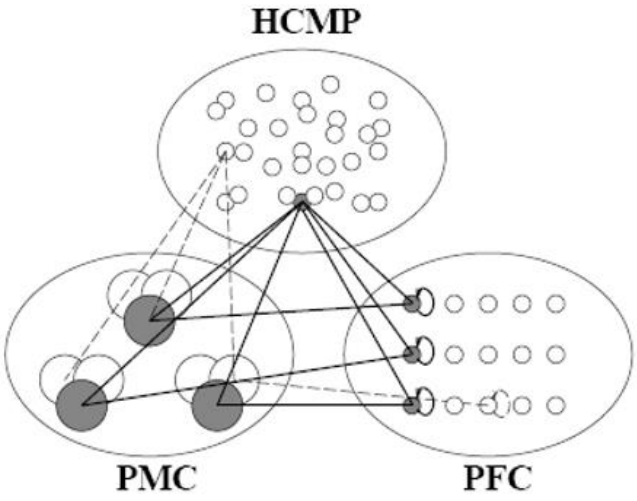


Diagram 6 Typy reprezentacji w modelu OBC.

Reprezentacje są symbolizowane przez zawarte w poszczególnych systemach kółka o mniejszej lub większej średnicy? Szare kółka oznaczają reprezentacje aktywowane na danym etapie realizacji celu/zadania. Tam, gdzie nachodzą na siebie kółka, mamy do czynienia z reprezentacjami rozproszonymi, co znaczy, że napływające informacje są w pewien sposób do nich dołączane i zaczynają tworzyć pewną neuronalną całość. Tam, gdzie są one odseparowane, mamy do czynienia z reprezentacjami wyizolowanymi i w dużym stopniu niezależnymi. Linie łączące poszczególne kółka oznaczają relacje istniejące między poszczególnymi składowymi całego układu. Linie ciągłe odnoszą się do relacji aktywnych w danym momencie, a przerywane do relacji istniejących, ale nieaktywnych. W przypadku PMC reprezentacje zgrupowane są wokół trzech modalności sensorycznych, z których każda realizuje swój specyficzny wzorzec przetwarzania (np. identyfikacja bodźca, realizacja programu motorycznego). Reprezentacje wykorzystywane przez PFC są wyizolowane i kombinatorycznie rozbite na poszczególne cechy (np. kolory lub kształty znaków w zadaniu AX-CT). Funkcjonujące w PFC połączenia rekurencyjne zapewniają stabilność (*robust*) reprezentacji w tym systemie. Mają one charakter samo-podtrzymujący (*self-maintenance*) w trakcie realizacji celu/zadania. Znaczy to, że nie wymagają one w tym czasie zewnętrznych bodźców, aby działać i są odporne na zaburzenia ze strony nieistotnych – z perspektywy zadania – informacji. W hipokampie (HCMP) związki między różnymi elementami przechowywane są w formie asocjacji obejmujących reprezentacje zawarte w PFC i PMC (np. jeśli pojawi się bodziec X i jest on zgodny z reprezentacją Z, to wykonaj ruch Y). Ponadto, HCMP stanowi rodzaj magazynu

pamięciowego dla stanów pośrednich, wykorzystywanych podczas realizacji zadania. Innymi słowy, omawiana struktura ma zdolność do szybkiego zapamiętywania sekwencji określonych reprezentacji. Zdolność ta jest z jednej strony użyteczna, z drugiej może prowadzić do trudności w określeniu statystycznie relewantnych zależności pośród zapamiętanych elementów. HCMP „stosuje" tzw. separację wzorców (*pattern separation*) w celu pokonania tej trudności. Elementy zaprezentowane w tym samym przedziale czasu łączone są w całości (epizody), niezależne od innych całości, nawet – jeśli dany epizod jest podobny do już wcześniej zapamiętanego. W ten sposób w HCMP pamiętany jest dodatkowy kontekst dla realizowanego aktualnie zadania (por. O'Reilly i in., 1999).

Systemy wskazane w modelu, a także funkcjonujące między nimi interakcje – wraz z nadzorowanymi przez nie reprezentacjami – stanowią bazę dla działania pamięci roboczej. Podobnie jak w podstawowej wersji hipotezy bramkowania dopaminergicznego, tak również w modelu pamięci roboczej istotnym elementem procesu realizacji zadania jest połączenie PFC-DA. Stabilizuje ono reprezentacje w PFC oraz umożliwia ich aktualizację na podstawie błędu predykcji nagrody. DA w tym układzie pełni zatem funkcję bramki, która blokuje lub odblokowuje dopływ informacji z otoczenia (PMC) oraz pamięci (HCMP). Warto w tym miejscu przypomnieć, że błąd predykcji nagrody jest jednym z kluczowych elementów algorytmu TDRL. To z jego pomocą kształtowana jest funkcja wartości, która po odpowiedniej liczbie interakcji agenta ze środowiskiem pozwala wyznaczyć optymalną strategię zachowań. Mamy zatem do czynienia, jak mawia się w żargonie informatycznym, z „reużyciem" tej samej informacji. Z jednej strony pozwala ona skutecznie pozyskiwać nagrody, a z drugiej stabilizuje, selekcjonuje i optymalizuje reprezentacje stanów świata niezbędnych do realizacji tego procesu.

Podwójna funkcja dopaminy, zdaniem Montague'a, jest kolejnym dowodem na to, że mózg, wbrew obiegowym opiniom, jest bardzo dobrze zaprojektowanym „systemem obliczeniowym". Uczeń Terrence'a Sejnowskiego uważa, że propozycję O'Reilly'a i innych badaczy można wzbogacić o jeszcze jedną ważną funkcję, mianowicie – o wykrywanie informacji odnoszących się do nagród „dostrzeżonych" w strumieniu danych sensorycznych. Tego typu zdolność jest podstawą, istotnego z perspektywy przetrwania, procesu decyzyjnego odpowiedzialnego za kontynuowanie lub zmianę aktualnie realizowanego celu. W zaproponowanym przez Montague'a ujęciu DA będzie reagować nie tylko na informacje dotyczące aktualnie realizowanego celu, ale również na nagrody lub obserwacje związane z zupełnie nowymi celami. Zidentyfikowana w ten sposób

możliwość nie musi zostać bezwarunkowo wykonana. Zanim dojdzie do ewentualnej aktywacji celu, najpierw zostanie on poddany ocenie: na ile związane z nim korzyści przewyższają te, które przyniesie realizacja bieżącego celu. Przedstawiona wizja mechanizmu odpowiedzialnego za zarządzanie celami ma, jak twierdzi Montague, swoje uzasadnienie empiryczne, w szczególności przekonujące są w jego opinii eksperymenty Donalda Stussa i Roberta Knighta (2002).

Wymienieni badacze, manipulując aktywnością wybranych obszarów PFC (poprzez zastosowanie określonych leków lub poprzez elektrostymulację), zidentyfikowali dwa charakterystyczne wzorce zachowań. Pierwszy z nich to tzw. perseweracja, czyli pełna i nieustanna koncentracja badanych na aktualnie realizowanym celu. Odpowiednia stymulacja powodowała, że uczestnicy eksperymentu nie byli w stanie przerwać jego realizacji nawet wówczas, gdy jego kontynuacja nie miała już żadnego sensu z perspektywy zakończenia zadania z sukcesem. Drugi wzorzec – skrajne rozproszenie – powodował, że badani nie byli w stanie dokończyć rozpoczętego zadania (Stuss & Knight, 2002). Zaobserwowane efekty wyraźnie pokazują, jak ważna jest równowaga w układzie aktywującym i dezaktywującym cele, którego zaburzenie może prowadzić do nieefektywnych wzorców zachowań.

Zaprezentowana powyżej hipoteza „nad-mocy" oraz hipoteza bramkowania dopaminergicznego – wraz z rozszerzeniami zaproponowanymi przez Reada Montague'a – stanowią próbę wyrażenia złożonych zachowań ludzkich za pomocą neuronalnych mechanizmów związanych z pojęciem nagrody oraz uczenia się ze wzmacnianiem.

Jednym z najważniejszych elementów przedstawionej konstrukcji jest niewątpliwie hipoteza dopaminergicznego błędu predykcji nagrody, za pomocą której wprowadzono do badań nad neurobiologicznymi podstawami zachowań niezwykle płodny eksplanacyjnie model oparty na algorytmie TDRL. Wymieniony algorytm nie tylko dobrze wyjaśnia przebieg określonych zjawisk neuronalnych (fluktuacji pobudzeń neuronów dopaminergicznych), ale również trafnie rekonstruuje określone schematy zachowań, w tym: zjawisko uczenia się pod wpływem wzmocnień (Schultz i in., 1997), efekt uzależnienia od określonych substancji (np. narkotyków) (Groman i in., 2019) czy sztywność w przypadku osób chorych na Parkinsona (Heisters, 2011). Wymienione przypadki można, zdaniem Reada Montague'a, znacząco rozszerzyć, jeśli przyjmiemy, zgodnie z hipotezą nad-mocy, że status nagrody uzyskają nie tylko zakodowane

genetycznie nagrody podstawowe (pożywienie, popęd seksualny, itp.), ale również określone idee, w tym tak abstrakcyjne – jak idea wolności politycznej. Możemy wyjaśnić, jak twierdzi Montague, po włączeniu tego typu idei w system pozyskiwania nagród (w szczególności z opcją cyklicznego załączania) nawet tak ekstremalne przypadki zachowań, jak zbiorowe samobójstwo członków sekty *Heaven's Gate* czy zaprzeczające instynktowi samozachowawczemu głodówki polityczne. Dane eksperymentalne (Gu i in., 2015) sugerują również, że mechanizm ewaluacji nagród odnosi się do szerszego zagadnienia, czyli do kontroli poznawczej. Na podobny związek wskazuje również hipoteza bramkowania dopaminowego, zgodnie z którą pozyskiwane ze środowiska informacje muszą być odpowiednio ocenione i skategoryzowane, zanim włączone zostaną w proces realizacji celu. W szczególności, istotne jest odróżnienie dystraktorów od informacji relewantnych dla danego celu, twierdzą O'Reilly, Braver i Cohen (1999). Ten sam mechanizm, zdaniem Montague'a, wpływa na proces wyższego rzędu, czyli na zarządzanie celami (patrz: tzw. rozszerzenie hipotezy bramkowania dopaminowego zaproponowane przez tego uczonego), za pomocą którego agent „decyduje": czy dany cel zostanie przerwany i zastąpiony innym celem, czy będzie kontynuowany.

Osiągnięcia podejścia obliczeniowego opartego w dużej mierze na hipotezie dopaminergicznego błędu predykcji nagrody są niewątpliwe. Czy zatem, trawestując tytuł książki Dennetta (*Consciousness explained*, 1991), możemy stwierdzić, że działania intencjonalne zostały wyjaśnione (*Intentional actions explained*)? Wydaje się, że na obecnym etapie badań takie stwierdzenie byłoby zdecydowanie na wyrost. Nietrudno zauważyć, że wspomniane koncepcje w dużym stopniu ignorują fakt, że stany intencjonalne są powiązane ze sobą i tworzą sieć, w której zawarta jest olbrzymia baza wiedzy wspomagająca realizację naszych zamierzeń i celów. Niewątpliwie, pomysł Montague'a, by rozszerzyć pojęcie nagrody i uznać, że określone idee (stany intencjonalne) również mogą uzyskać tego typu status, pozwala wyjaśnić wiele zachowań traktowanych jako nieracjonalne, a w szczególności autodestrukcyjne (np. można w ten sposób wyjaśnić zachowania „ekstremalne", będące skutkiem skrajnego „przewartościowania" określonego celu/idei/ideologii). Można jednak mieć wątpliwości, czy tego typu podejście jest wystarczające, aby wyjaśnić zjawiska takie jak: świadome planowanie czy modyfikowanie własnych zachowań pod wpływem informacji zewnętrznych, bez ich związku z jakimkolwiek wzmocnieniem. Aby pokazać, że wskazane wątpliwości są uzasadnione,

warto przyjrzeć się pracom teoretyków i praktyków uczenia maszynowego, którzy w najróżniejszych dziedzinach wykorzystują algorytmy RL i dobrze znają jego ograniczenia.

3.4 Wybrane rozszerzenia metody uczenia się ze wzmacnianiem

Warto, zanim przedstawione zostaną bardziej szczegółowe analizy dotyczące poszczególnych rozszerzeń metody uczenia się ze wzmacnianiem, doprecyzować relację między samą metodą uczenia się ze wzmacnianiem a jej rozmaitymi rozszerzeniami. Paweł Cichosz twierdzi, że metoda uczenia się ze wzmacnianiem stanowi obecnie niezależny paradygmat w obszarze uczenia maszynowego (oprócz niej zwykle wyróżnia się jeszcze dwa inne paradygmaty: (1) uczenie z nadzorem oraz (2) uczenie bez nadzoru). W paradygmacie uczenia się ze wzmacnianiem uczeń ma się nauczyć „celowego zachowania na podstawie dynamicznych interakcji ze środowiskiem. Interakcje te przybierają postać obserwowania przez ucznia stanów środowiska, wykonywania akcji i obserwowania oceniających efekty tych akcji rzeczywistoliczbowych nagród, nazywanych też wartościami wzmocnienia.” (Cichosz, 2007, s. 712). Ten ogólny cel można zrealizować na wiele sposobów, dlatego istnieje zbiór algorytmów należących do tego paradygmatu. Do najbardziej popularnych należą: TD-λ (gdy λ=0, otrzymujemy algorytm TDRL), Q-learning, AHC, SARSA. Wymienione algorytmy realizują ten sam cel: poszukują optymalnej strategii zachowań dla danego środowiska. Niestety, znalezienie optymalnej strategii wymaga realizacji kosztownego i czasochłonnego procesu eksploracji środowiska. Poszczególne algorytmy stosują w tym kontekście różne techniki, jednak każda z nich ograniczona jest przez bezpośrednio dostępne obserwacje (patrz: tzw. własność Markova). Wiedza ogólna o danym środowisku, np. o obecnych w nim prawidłowościach, nie jest wykorzystywana w tego typu algorytmach. Znaczy to, że agent korzystający z tej metody, zanim znajdzie optymalną strategię doboru zachowań, musi poprzez interakcje zbudować wiarygodny statystycznie model środowiska (tzw. MDP – *Markov Decision Process*) – w szczególności: zidentyfikować nagradzające stany oraz prowadzące do nich zachowania. Taki proces potrafi być kosztowny i długotrwały. W związku z tym poszukiwane są sposoby skrócenia eksploracji poprzez zastosowanie wiedzy wykraczającej poza bezpośrednią obserwację (Grześ, 2010). Jedno z podejść polega na „wstrzyknięciu” wiedzy symbolicznej w algorytm, wiedzy, która pozwoli agentowi ocenić, czy w danym obszarze dalsza eksploracja środowiska ma sens. Jedno z zaprezentowanych poniżej rozszerzeń metody uczenia się ze wzmacnianiem dotyczy tego właśnie zagadnienia.

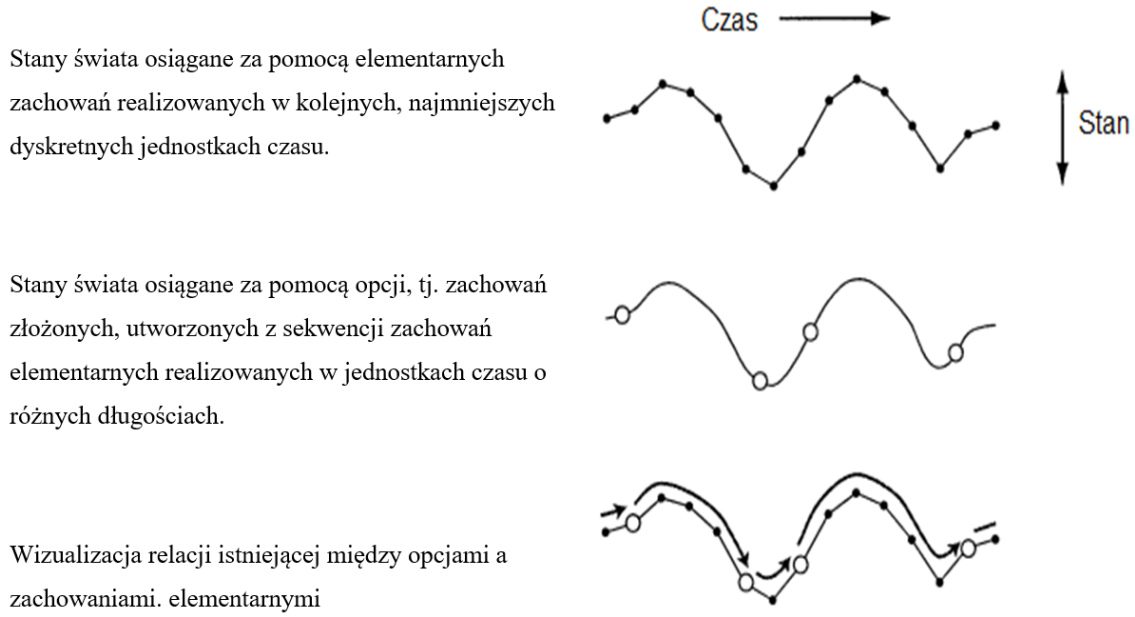
Brak efektywnego mechanizmu planowania działań jest kolejnym problemem, z którym zmagają się badacze pracujący w ramach paradygmatu uczenia się ze wzmacnianiem. Jego rozwiązanie wymaga, zdaniem Richarda Suttona, uzupełnienia algorytmu RL o zdolność do łączenia elementarnych zachowań w większe jednostki oraz takiego ich konceptualizowania, by możliwe było posłużenie się nimi podczas konstruowania planu (w „żargonie" uczenia maszynowego mówi się o tzw. problemie reprezentowania wiedzy na wielu poziomach czasowej abstrakcji) (Sutton, Precup, & Singh, 1999). Łatwo dostrzec, że odniesieniem dla tego typu badań są ludzkie działania intencjonalne, które cechują się rozbudowaną hierarchią oraz kompozycyjnością wspomaganą mechanizmami planowania. W dalszej części niniejszego rozdziału omówione zostaną zasygnalizowane powyżej rozszerzenia algorytmu.

Hierarchiczne uczenie się ze wzmacnianiem

Podstawowa wersja metody uczenia się ze wzmacnianiem pozwala, jak już wspomniano, realizować założony cel tylko za pomocą działań elementarnych, np. „idź w przód", „idź w tył", itp. Z perspektywy ludzkich działań znaczyłoby to, że tego typu metoda mogłaby co najwyżej wyjaśnić bardzo wąską grupę zachowań. Trudno wyobrazić sobie, by świadome planowanie odbywało się na poziomie prostych ruchów w rodzaju: „przesunąć lewą nogę do przodu", „przesunąć prawą nogę do przodu" (etap podejścia do drzwi), „podnieść prawą rękę do poziomu klamki", „chwycić prawą ręką za klamkę", „nacisnąć prawą ręką klamkę", „pociągnąć prawą rękę do siebie" (etap otwarcia drzwi), „przesunąć lewą nogę do przodu", „przesunąć prawą nogę do przodu" (etap przejścia przez drzwi), itd. Nawet tak prosty przykład wyraźnie pokazuje, jak ważne jest, by podmiot realizujący działania potrafił łączyć sekwencje prostych ruchów w jednostki wyższego rzędu i pracować na tych jednostkach, tworząc z nich złożone działania, a więc jednostki jeszcze wyższego rzędu. Podejście takie zapewnia zachowanie ustrukturowanej, wielopoziomowej hierarchii zachowań, która identyfikowana jest jako działanie złożone. Jednym z ciekawszych rozszerzeń metody RL, odnoszącym się do tego problemu, jest opracowana przez zespół Richarda Suttona idea hierarchicznego uczenia się ze wzmacnianiem. Kluczowe rozszerzenie polega na wprowadzeniu do algorytmu pojęcia opcji (*option*). Opcja to nic innego, jak uogólnienie pojęcia zachowania (tzw. akcji). Może ona reprezentować zarówno zachowanie elementarne, jak i całą sekwencję tego rodzaju zachowań (np. „otwarcie drzwi", „przejście przez korytarz", „zadokowanie robota").

Każdą opcję charakteryzują trzy elementy: (1) stan początkowy, (2) strategia sterująca doborem zachowań elementarnych w ramach opcji, (3) stany końcowe. Istotne okazuje się to, że strategia zdefiniowana z wykorzystaniem opcji nie zawsze będzie równie optymalna jak strategia globalna wyznaczona wyłącznie na podstawie zachowań elementarnych. Jest to konsekwencja, którą się ponosi z powodu planowania, a co za tym idzie – skrócenia czasu eksploracji środowiska. Ma to niebagatelne znaczenie dla tzw. skalowalności algorytmu, czyli możliwości jego wykorzystania do odpowiednio złożonych problemów[43].

Poniższy diagram prezentuje ideę funkcjonowania opcji:



Rys. 1. Opcje w hierarchicznym uczeniu się ze wzmacnianiem.

Metoda hierarchicznego uczenia się, oprócz możliwości reagowania na suboptymalne efekty działania opcji (patrz: mechanizm kończenia), posiada również zdolność (w tzw. trybie *off-policy*) do optymalizowania strategii przypisanych do opcji. Sekwencja zachowań składająca się na daną opcję może uzyskać w ten sposób bardziej efektywną postać. Z perspektywy ludzkich zachowań tego typu możliwość to nic innego, jak pewna umiejętność, którą podmiot działający może wykorzystywać w wielu kontekstach, do realizacji różnych celów. Przykładem zoptymalizowanej umiejętności mogą być takie działania jak: „otwieranie drzwi", „wchodzenie po schodach", „rzucanie różnego rodzaju

[43] Warto dodać, że problem suboptymalności opcji można do pewnego stopnia zniwelować, wprowadzając mechanizm tzw. kończenia(*termination*), polegający na przełączaniu się między strategią wykorzystującą opcje a strategią globalną, która opiera się na zachowaniach elementarnych. Strategie przypisane do opcji ograniczają eksplorację, ale nie blokują możliwości zastosowania podejścia bazowego, tj. pełnej eksploracji środowiska za pomocą zachowań elementarnych.

przedmiotami", itp. Wymienione przykłady zwykło się w psychologii poznawczej klasyfikować jako tzw. wiedzę proceduralną – „wiedzę-jak", która pozwala sprawnie realizować złożone działania, ale którą trudno jest wyeksplikować. Pojęcie opcji łatwo jest też zinterpretować jako dyspozycję/umiejętność tła (Searle, 1983, s. 143) lub działanie podstawowe, czyli w kategoriach teorii intencjonalności Johna Searle'a. Warto przypomnieć, że tego typu umiejętności i działania pełnią fundamentalną rolę zarówno z perspektywy kontroli zachowań, jak i budowania sieci stanów intencjonalnych. Związek hierarchicznego uczenia się z umiejętnościami tła zostanie omówiony dokładniej w ostatnim rozdziale pracy.

Polepszenie procesu eksploracji w metodzie uczenia się ze wzmacnianiem poprzez zastosowanie wiedzy dziedzinowej

Drugim rozszerzeniem metody uczenia się ze wzmacnianiem, o którym warto wspomnieć w kontekście badań nad działaniami intencjonalnymi, jest optymalizacja procesu eksploracji środowiska poprzez zastosowanie wiedzy dziedzinowej. Czasochłonność i kosztowność procesu eksploracji, jak już wspomniano, jest jednym z największych problemów metody uczenia się ze wzmacnianiem. W praktycznych zastosowaniach, w których przestrzeń stanów świata jest liczna (czasami nawet ciągła), niemal bezużyteczna okazuje się standardowa wersja algorytmu. W takim środowisku, nawet przy zastosowaniu dużej mocy obliczeniowej komputera, proces uczenia się trwałby dziesiątki lat i nadal jego rezultaty byłyby mizerne. Współcześnie problem ten rozwiązuje się na kilka sposobów. Jednym z nich jest zastosowanie funkcji aproksymujących, które pozwalają – na bazie próbki stanów środowiska – wyznaczyć przybliżoną postać funkcji wartości V dla dowolnego stanu. Przykładem tego typu aproksymatorów mogą być sztuczne sieci neuronowe lub liniowe kombinacje cech. Wskazane metody pozwalają w istotny sposób zredukować zakres przeszukiwania przestrzeni stanów świata. Tego typu podejście umożliwia zastosowanie metod uczenia się ze wzmacnianiem do tak złożonych problemów, jak gra w trick-tracka (*Backgammon*) – liczba stanów=10^{20}, jak gra w Go – liczba stanów = 10^{170}, a nawet nawigowanie helikopterem – liczba stanów = przestrzeń ciągła (Silver i in., 2016). Możliwość zastosowania aproksymatorów funkcji wartości stanowi niewątpliwie znaczący krok w rozwoju metody uczenia się ze wzmacnianiem. Obrazowo można powiedzieć, że wprowadzenie tego typu funkcji znacząco poszerzyło model świata. Od tej pory wystarczy, że agent zidentyfikuje wybrane cechy danego

środowiska, aby na tej podstawie prawidłowo wyznaczać funkcję wartości, a w konsekwencji – zdecydować o dalszym kierunku działania. Należy równocześnie zauważyć, że wskazane rozszerzenie bazuje na takiej samej informacji, co podstawowa wersja algorytmu, tj. na sygnałach nagrody generowanych przez środowisko oraz obserwacjach. Podstawowa wersja algorytmu nie przewiduje wykorzystania innego rodzaju wiedzy do eksploracji środowiska, w którym funkcjonuje agent, niż wymienione typy. Z perspektywy modelowania ludzkich działań intencjonalnych takie ograniczenie wydaje się być istotnym mankamentem. Wskazać bowiem można wiele przykładów na to, że nasze zachowania modyfikowane są pod wpływem informacji całkowicie zewnętrznej, uzyskanej na podstawie obserwacji pochodzącej od innych osób lub zaczerpniętej z wiedzy z nauk ścisłych czy humanistycznych. Przykładami, które dobrze obrazują „zysk" z włączenia zewnętrznych źródeł wiedzy w proces eksploracji środowiska, mogą być różnego rodzaju przewodniki opisujące niebezpieczne miejsca lub zjawiska na kuli ziemskiej. Łatwo wyobrazić sobie, jaka byłaby skala wypadków podczas wypraw w Himalaje, gdyby każdy alpinista musiał samodzielnie odkrywać rządzące tym środowiskiem prawa i reguły. Jednym ze sposobów włączenia dodatkowej wiedzy w metodę uczenia się ze wzmacnianiem jest zastosowanie „nagród kształtujących" (*reward shaping*).

> *Nagroda kształtująca nie pochodzi ze środowiska. Reprezentuje ona dodatkową informację, którą projektant włącza w system na bazie swojej uprzedniej wiedzy opracowanej na podstawie znajomości problemu.* (Grześ, 2010, s. 34).

Formalnie nagrody kształtujące reprezentuje się w postaci funkcji $\mathbf{F(s_t, s_{t+1})}$, która określa wartość nagrody kształtującej dla dwóch kolejnych stanów świata. Po uwzględnieniu tego rozszerzenia wartość błędu predykcji obliczana jest w następujący sposób:

$$\Delta = r_0 + \mathbf{F(s_t, s_{t+1})} + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t);$$

Zgodnie z powyższą formułą, jeśli $F(s_t, s_{t+1})$ będzie wynosiło 0 dla dowolnej pary s_t, s_{t+1}, to uzyskamy standardowy błąd predykcji nagrody. Dobrym przykładem pokazującym, w jaki sposób nagrody kształtujące mogą pozytywnie wpływać na przyspieszenie procesu uczenia się, może być zadanie nawigacyjne[44], opisane funkcją F:

[44] Zadania nawigacyjne polegają na nauczeniu się przez robota znajdowania w danym środowisku najkrótszej drogi do celu.

$$F(s_t, s_{t+1})\ s_{t+1}) \begin{cases} 0{,}5, \text{gdy } |s_d - s_{t+1}| < |s_d - s_t| \\ 0, \text{gdy } |s_d - s_{t+1}| >= |s_d - s_t| \end{cases}$$

Powyższa funkcja nagradza agenta, gdy ten wykona akcję przybliżającą go do stanu docelowego s_d. Warto podkreślić, że funkcja F nie jest uniwersalna. Jej postać może różnić się nie tylko w zależności od rodzaju problemu, ale również od wiedzy i doświadczenia projektanta. W określonych przypadkach jej zastosowanie, wbrew intencjom twórcy, może prowadzić do wydłużenia procesu eksploracji, a w skrajnych sytuacjach – do niezamierzonego przedefiniowania problemu w sposób, który uniemożliwi jego rozwiązanie. W 1999 roku, Adrew Ng udowodnił, że funkcja F musi posiadać określoną postać, by nie zaburzyła kluczowych własności algorytmu RL (Grześ, 2010, s. 35).

Z przedstawionego przykładu wnosić można, że algorytm uczenia się ze wzmacnianiem może zostać w określony sposób poszerzony o wiedzę dziedzinową. Pojawia się jednak pytanie: jakiego rodzaju wiedza może być w ten sposób wyrażona? Na obecnym etapie badań możemy stwierdzić jedynie, że nagrody kształtujące reprezentują efektywnie wyłącznie wiedzę proceduralną (Grześ, 2010, s. 39), tj. wiedzę, którą sztuczne systemy wykorzystują podczas planowania lub realizacji działań. Przykładem mogą być instrukcje zapisane w języku STRIPS, który wykorzystywany jest do automatycznego planowania. W formie symbolicznej w tego typu językach – dla abstrakcyjnie zdefiniowanego środowiska (np. zbioru pomieszczeń w budynku) – można wyrazić:

- stany początkowe (np. robot znajduje się w korytarzu),
- tzw. „sztywne" fakty opisujące strukturę środowiska (np. dwa pomieszczenia sąsiadują ze sobą; istnieje przejście pomiędzy pomieszczeniami X i Y; liczba nagród w pomieszczeniu A wynosi r_A),
- dopuszczalne zachowania (tzw. operatory) (np. możliwość przemieszczenia się pomiędzy pomieszczeniami, możliwość pobrania nagrody),
- stan docelowy (np. dotarcie do pokoju X i zebranie wszystkich dostępnych w całym mieszkaniu nagród).

Wbudowane w interpreter języka STRIPS mechanizmy wnioskowania pozwalają, na podstawie zadanego skryptu, wydedukować plan zachowań, który pozwoli robotowi przejść ze stanu początkowego do stanu docelowego. Tak opracowany plan, czyli *de facto*

odpowiednio zdefiniowana sekwencja stanów, może następnie zostać wykorzystany do wyznaczenia wartości nagród kształtujących, które, jak już wspomniano, stanowią interfejs pomiędzy algorytmem RL a wiedzą dodatkową, zewnętrzną w stosunku do tej, którą agent pozyska w przyszłości bezpośrednio ze środowiska. W wysokopoziomowym planie wartość nagrody kształtującej określa się w następujący sposób: najpierw wyznacza się funkcję między stanami niskopoziomowymi a stanami abstrakcyjnymi, na których operuje plan. Następnie definiuje się funkcję potencjału F, która jako argument przyjmuje stan abstrakcyjny, a pozycję w planie jako wartość. Innymi słowy, im dany stan abstrakcyjny jest bliżej stanu docelowego, tym ma wyższą wartość. Agent, eksplorując środowisko i dysponując tego typu funkcjami, zdecydowanie bardziej preferuje miejsca (pomieszczenia), które znalazły się w planie, niż te, które są nieistotne z perspektywy planu. Co ciekawe, algorytm wykorzystujący nagrody kształtujące pozwala agentowi nauczyć się optymalnej polityki zachowań nawet wówczas, gdy plan jest wadliwy. Oczywiście, w takim przypadku proces uczenia znacząco się wydłuży.

Opisana zasada działania nagród kształtujących wyraźnie wskazuje na ważną z perspektywy działań intencjonalnych możliwość: mechanizm uczenia się ze wzmacnianiem może zostać „wsparty" przez wysokopoziomowy plan wyrażony w formie wiedzy symbolicznej (np. w formie skryptu języka STRIPS). Można sformułować ciekawą teoretycznie hipotezę, przyjmując, że istnieje ścisły związek między wiedzą symboliczną a stanami intencjonalnymi, a mianowicie: dwie niezależne formy reprezentowania wiedzy o świecie – (1) informacje dostarczane w formie sygnałów nagrody i pobudzeń sensorycznych oraz (2) wiedzę symboliczną np. reguły logiczne – można włączyć w jeden złożony system działający zgodnie z regułami stosowanymi w algorytmach uczenia się ze wzmacnianiem. W ostatnim rozdziale pracy wskazana hipoteza zostanie wykorzystana do modelowania wybranych aspektów działań intencjonalnych.

Podsumowanie

„Czego nie potrafię stworzyć, tego nie potrafię zrozumieć." – twierdził Richard Feynman (*Richard Feynman cytaty*, 2016). To stwierdzenie wyraźnie pokazuje, że zrozumiemy jakieś zjawisko dopiero wtedy, kiedy potrafimy skonstruować jego model. Neuronaukowcy zaczynają formułować podobne wymagania (por. Montague, 2006, ss. 142–146). Przedstawiona w niniejszym rozdziale hipoteza dopaminergicznego błędu predykcji

nagrody jest przykładem na to, jak płodne poznawczo może być połączenie podejścia obliczeniowego z neurobiologicznym. Wskazana hipoteza nie tylko wyjaśnia przebieg fluktuacji neuronów dopaminergicznych, ale również tłumaczy obserwowane reakcje behawioralne agenta. Algorytm, leżący u jej podstaw, objaśnia zarówno przypadki standardowe (uczenie warunkowe), jak i zaburzenia wynikające z destabilizacji mechanizmu odpowiedzialnego za wyznaczanie błędu predykcji nagrody (np. problem uzależnień; sztywność w chorobie Parkinsona). Formalna definicja algorytmu TDRL pozwala nie tylko na jego implementację, ale również na analizę jego złożoności oraz na określenie warunków, w których będzie on efektywny. W ten sposób udało się zidentyfikować główne problemy tego typu uczenia się. Wśród nich szczególnie istotny okazuje się dylemat: eksploracja vs. eksploatacja. By go rozwiązać, do algorytmów uczenia się ze wzmacnianiem wprowadza się specjalne rozszerzenia, które minimalizują problem długiego czasu eksploracji. Wskazane cechy i problemy, związane z algorytmem uczenia się ze wzmacnianiem, staną się przedmiotem szczegółowej analizy w ostatnim rozdziale pracy, który poświęcony będzie zintegrowanemu modelowi działań intencjonalnych. W zaproponowanym rozwiązaniu podsystem implementujący ten typ uczenia się będzie pełnił kluczową rolę w układzie odpowiedzialnym za kontrolę zachowań. W najbardziej zaawansowanej wersji w model działań intencjonalnych włączone zostaną wszystkie wskazane w niniejszym rozdziale rozszerzenia algorytmów RL, by za ich pomocą zrekonstruować charakterystyczne cechy działań intencjonalnych, w szczególności: zdolność do planowania działań opartych na wiedzy zawartej w sieci stanów intencjonalnych.

Analizy przeprowadzone w rozdziałach drugim i trzecim, które dotyczą wpływu stanów intencjonalnych na przebieg zachowań celowych, opisane zostały przy wykorzystaniu dwóch zasadniczo odmiennych aparatów pojęciowych. Wydaje się jednak, że obydwa podejścia mają jedną wspólną cechę – występujące w nich reprezentacje traktuje się jako zasadniczo adekwatne odwzorowania rzeczywistości. W obu ujęciach zwraca się uwagę na to, że efektywne działania wymagają adekwatnego modelu środowiska, zarówno w jego wymiarze fizycznym, jak i społecznym. Potwierdzeniem takiego stanu rzeczy są różnego rodzaju halucynacje, urojenia, konfabulacje i inne formy zaburzeń zniekształcające sposób reprezentowania rzeczywistości i prowadzące często do nieefektywnych zachowań, które w skrajnych przypadkach powodują przedwczesną śmierć. Wydaje się jednak, że problem błędnego reprezentowania rzeczywistości nie dotyczy tylko osób chorych. Wielu z nas na co dzień doświadcza różnego rodzaju zjawisk, które albo w ogóle nie występują w świecie

realnym (omamy, np. słyszenie głosów albo urojenia jak np. przekonanie o celowym wywołaniu epidemii), albo – ze względu na specyficzne okoliczności – są deformacjami zdarzeń ze świata realnego. Nie do końca wiemy, dlaczego w danej sytuacji wybraliśmy tak, a nie inaczej lub dlaczego powiedzieliśmy coś, czego później żałowaliśmy. We wszystkich tego typu sytuacjach ujawnia się pewna niepokojąca cecha umysłu ludzkiego, a mianowicie – jego zdolność do konstruowania wyjaśnień i form reprezentowania rzeczywistości, które z samą rzeczywistością mogą mieć mało wspólnego. Na obecnym etapie badań trudno jest nam rozpoznać, jak funkcjonują wymienione sposoby reprezentowania świata. Niełatwo też jest wskazać mechanizmy, które decydują o tym, że w określonych dziedzinach życia sieć stanów intencjonalnych jest dobrze ugruntowana i adekwatnie odnosi się do rzeczywistości, a w innych dziedzinach jest podatna na złudzenia i prezentuje zdeformowany obraz świata. Dodatkowo, sytuację komplikuje holistyczny charakter treści poszczególnych stanów intencjonalnych oraz ich tylko częściowa dostępność w polu świadomości. Oczywiście, cały ten kontekst ma również wpływ na nasze intencje, a co za tym idzie – na nasze działania. Od wielu lat prowadzone są badania, w których próbuje się zidentyfikować procesy i mechanizmy wpływające na kształt intencji oraz określić jej funkcję. Doniosłymi osiągnięciami w tym zakresie może pochwalić się psychologia intencji, subdyscyplina neuropsychologii, która w systematyczny sposób bada strukturę i funkcję tego szczególnego stanu intencjonalnego. Najważniejsze osiągnięcia oraz wnioski płynące z prowadzonych w tym zakresie badań będą przedmiotem analizy w kolejnym rozdziale.

4 Korelacyjno-interpretacyjny status stanów intencjonalnych towarzyszących prostym działaniom intencjonalnym

Wyobraźmy[45] sobie, że znajdujemy się w Hanowerze na największych targach automatyki przemysłowej w Europie. Podczas zwiedzania dostrzegamy stoisko firmy oferującej nietypowe automaty do wydawania kawy. Zaciekawieni, ponieważ z zewnątrz urządzenie przypomina kabinę do robienia fotografii paszportowych, postanawiamy przetestować nowość. Słyszymy, po wejściu do kabiny, jak głos w tle prosi nas o zajęcie odpowiedniej pozycji na wprost ścianki z wyeksponowanymi zdjęciami poszczególnych rodzajów kawy. Co ciekawe, przy żadnym z nich nie ma przycisku, który pozwalałby dokonać wyboru. Po chwili słyszymy następne polecenie: „Proszę wybrać rodzaj kawy". W tym samym momencie do naszej głowy zbliża się miniaturowy skaner mózgu w formie hełmu stosowanego w grach wirtualnych. Zaczynamy, nieco zmieszani tym technologicznym wyposażeniem, przyglądać się ofercie. Po przeprowadzonej przez nas analizie wszystkich zdjęć pojawia się tradycyjne wahanie: na co się zdecydować. Wybór jest bogaty. Gdy ciągle się jeszcze zastanawiamy, nagle słyszymy, że automat zaczyna coś przygotowywać. Po chwili już wiemy, że największą ochotę mielibyśmy na *cappuccino* i kiedy już zamierzamy powiedzieć *'cappuccino'* przeżywamy mały szok. Głos wydobywający się z głośnika prosi nas, abyśmy odebrali wybraną przez nas kawę z podajnika umieszczonego po prawej stronie w kabinie. Nasze zdumienie staje się jeszcze większe, gdy podnosząc kubek, zauważamy na nim naklejkę z informacją: czas faktycznego podjęcia decyzji (zmierzony na podstawie zarejestrowanej aktywności mózgu) – 13:00:00.0, czas uświadomienia decyzji – (zmierzony moment powzięcia zamiaru) 13:00:01.0.

[45] Obszerne fragmenty niniejszego rozdziału zostały opublikowane w artykule: (Marcin Cichosz, 2010).

Przywołany przykład pozwala wyobrazić sobie, do jakich zaskakujących efektów mogą prowadzić wyniki współczesnych badań nad ludzkimi aktami wolicjonalnymi. Obecnie nie dysponujemy jeszcze narzędziami, które pozwalałyby w praktyce zrealizować powyższy scenariusz, ale – jak się wydaje – jest to raczej kwestia czasu i stopnia zaawansowania rozwoju technologicznego, niż jakichś fundamentalnych ograniczeń.

Inspiracją dla przedstawionej innowacji są dwa niezwykle intrygujące eksperymenty. Pierwszy, przeprowadzony w 1963 roku przez Williama Greya Waltera, pokazał następujący efekt: osoby, którym bezpośrednio podłączono pod korę motoryczną czujnik reagujący na nagły wzrost aktywności (*bursts of recorded activity*), a ten z kolei sprzężono z mechanizmem sterującym przerzucaniem slajdów w projektorze, doświadczali dziwnego uczucia. Tuż przed tym, kiedy w ich umysłach pojawiała się intencja, by – poprzez naciśnięcie przycisku atrapy – przejść do następnego slajdu, projektor był już w trakcie wykonywania oczekiwanej zmiany. Innymi słowy, moment uświadomienia sobie chęci zmiany slajdu był opóźniony względem zwiększonej aktywności kory motorycznej. Badani odnosili wrażenie, że projektor potrafił przewidzieć ich decyzję:

> *[Badani] raportowali, że tuż przed tym, jak zamierzali nacisnąć przycisk, zanim faktycznie zdecydowali się to zrobić, projektor zmieniał slajd – wskazany efekt wywoływał wrażenie, że naciskając przycisk spowodują, iż projektor zamiast o jeden, przejdzie o dwa slajdy do przodu![46]*

Podobny, choć znacznie precyzyjniejszy wynik, uzyskał w 1983 roku zespół kierowany przez Benjamina Libeta. W opracowanym przez amerykańskiego neurofizjologa eksperymencie z użyciem EEG udało się po raz pierwszy zmierzyć czasowy przebieg prostego aktu wolicjonalnego polegającego na zgięciu palca w swobodnie wybranej przez badanego chwili (Libet i in., 1983b). Libet, podobnie jak Walter, zaobserwował różnicę między momentem aktywacji struktur mózgu zaangażowanych w realizację ruchu a opóźnionym momentem, w którym badany świadomie decydował, że chce zgiąć palec. Obydwa eksperymenty dobrze pokazują, jak z pozoru prosty akt wolicjonalny jest tak naprawdę rezultatem współpracy szeregu nieświadomych i świadomych procesów angażujących wiele struktur mózgowych. Od lat prowadzone są badania zmierzające do

[46] "They reported that just as they were "about to" push the button, but before they had actually decided to do so, the projector would advance the slide - and they would find themselves pressing the button with the worry that it was going to advance the slide twice!" (D. Dennett & Kinsbourne, 1992, s. 199).

doprecyzowania przebiegu tego typu aktów, nazywanych w psychologii: „działaniami dowolnymi" lub „działaniami intencjonalnymi". Od pozostałych typów zachowań odróżnia je (Haggard, 2005): (1) względna niezależność od bodźców zewnętrznych[47], (2) związek z pewnym zamiarem oraz (3) towarzyszący im zbiór asocjacji, nabyty w procesie uczenia. Ponadto, działania tego typu na ogół poprzedzone są procesami planowania i rozumowania, a ich wykonanie wymaga odpowiednio skupionej uwagi. Przeciwieństwem działań intencjonalnych są odruchy, które zawsze silnie powiązane są z określonymi bodźcami i w zasadzie całkowicie znajdują się poza świadomą kontrolą. Odruchy mają charakter wrodzony – zawdzięczamy je zatem określonej ścieżce ewolucyjnej. Działania dowolne natomiast wykształcane są przez agenta stopniowo w wyniku procesów uczenia i ciągłego monitorowania uzyskiwanych rezultatów. Na skutek tego, nasze funkcjonowanie w środowisku staje się coraz skuteczniejsze. Ta pobieżna charakterystyka wydobywa na jaw jedną z kluczowych cech działań intencjonalnych, mianowicie: towarzyszące im poczucie świadomej kontroli. Gdy dokądś zmierzamy, realizujemy jakiś plan, mamy silne poczucie, że bez świadomie wybranego i podtrzymywanego w trakcie działania celu, bez świadomej kontroli kolejnych etapów działania w świecie nie nastąpiłaby zaprojektowana przez nas zmiana. Jeśli w trybie introspekcji, rozumianej jako „obserwacja" własnych stanów umysłowych, odwołamy się do raportów dotyczących konstruowanych przez nas planów i zamiarów, to w zasadzie nie mamy wątpliwości, że tego typu działania umysłowe towarzyszące działaniom fizycznym są inicjowane, wykonywane i kontrolowane przez świadomy swoich pragnień i wyborów podmiot. Wytworzonego na podstawie tych doświadczeń przekonania nie były i nie są w stanie podważyć argumenty filozoficzne odwołujące się do postulowanego przez naukę niemal-determinizmu[48] (Honderich, 2001). Jednak, jeśli uznamy wyniki badań przeprowadzonych przez Benjamina Libeta, to stajemy przed trudnym dylematem: czy zaufać danym uzyskanym w trybie introspekcji, czy uznać nadrzędność obiektywnych danych neurofizjologicznych? Najczęściej te ostatnie dane interpretuje się jako dowód na epifenomenalny status intencji w działaniu, co z kolei prowadzi do przyjęcia tzw. iluzyjnej koncepcji świadomej woli. Uważam, że wskazany wyżej problem jest o wiele bardziej złożony i wymaga wprowadzenia kilku istotnych dystynkcji, zanim dokona się wyboru między koncepcją intencjonalności ufundowaną na

[47] Znaczy to, że działanie jest efektem autonomicznej, swobodnie podjętej decyzji agenta, czyli nie jest bezpośrednią, niezwłoczną reakcją na docierający z otoczenia bodziec.

[48] Niemal-determinizm w ujęciu Teda Hondricha obejmuje (1) indeterminizm kwantowy oraz determinizm obecny w strukturach makroskopowych (Honderich, 2001).

subiektywnym poczuciu sprawstwa a koncepcją traktującą intencjonalność jako przejaw działania obiektywnych procesów mózgowych. Głównym celem tego rozdziału jest szczegółowa rekonstrukcja fenomenów składających się na przebieg prostego działania intencjonalnego, aby na ich podstawie można było skonstruować zintegrowany model działania intencjonalnego.

Rozdział podzielony został na trzy części. W pierwszej omówione będą dane eksperymentalne odnoszące się do dwóch głównych składowych **intencji w działaniu**, tj. do (1) poczucia chęci działania (*sens of urge*) oraz (2) odniesienia do docelowego obiektu lub zdarzenia (*reference forward to the goal object or event*). W kolejnej części zaprezentowany zostanie drugi – obok intencji w działaniu – istotny składnik działania intencjonalnego, czyli **poczucie sprawstwa**, niezwykle ciekawy fenomen, którego interpretację psychologiczną opracował Daniel Wegner. W ostatniej części tego rozdziału pokażę, jak każde z obydwu wymienionych wyżej zjawisk odnosi się do idei leżących u podstaw koncepcji przyczynowości intencjonalnej wprowadzonej przez Johna Searle'a.

4.1 Intencja w działaniu w ujęciu psychologii intencji

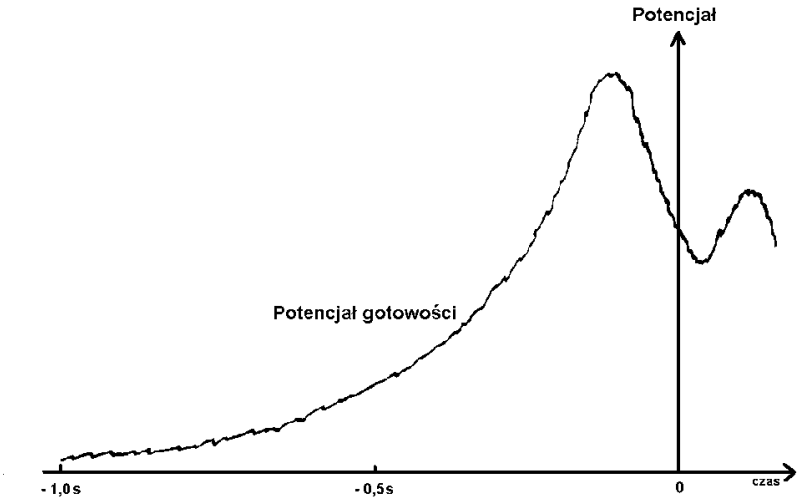
John Searle podczas wystąpienia w ramach konferencji TED, żartując z badaczy kwestionujących wpływ świadomości na nasze zachowania, stwierdził:

> *Obiecałem opowiedzieć o kilku niedorzecznościach mówionych o świadomości. Po pierwsze, świadomość nie istnieje. Jest iluzją, jak zachód słońca. Nauka wykazała, że zachody słońca i tęcze są iluzją, zatem świadomość jest iluzją. Po drugie, może świadomość istnieje, lecz jest czymś innym – programem komputerowym w mózgu. Po trzecie, naprawdę istnieje tylko zachowanie (niesamowite, jak wpływowy był behawioryzm). Wrócę do tego. Po czwarte, może świadomość istnieje, ale dla świata to bez znaczenia. Jak duchowość mogłaby tu cokolwiek poruszyć? Kiedy tak mówią, myślę: «Chcesz zobaczyć, jak dusza coś poruszy?» Spójrz! Świadomie decyduję podnieść rękę i ta cholerna ręka się unosi. (Śmiech). Zauważcie jedno. Nie mówimy: «To jak z pogodą w Genewie, czasami się poprawia, czasami nie». Nie, psiakrew! Poruszam nią, kiedy chcę* (Searle, 2013).

O ile trzy pierwsze uwagi mają charakter głównie filozoficzny, o tyle czwarta bezpośrednio odnosi się do naszych najbardziej podstawowych doświadczeń. Wskazane przez Searle'a

powiązanie świadomej intencji z ruchem ręki zdawałoby się, oczywiste, stało się przedmiotem nie tylko namysłu filozoficznego, ale również badania neurofizjologicznego.

Badaczami, którzy jako pierwsi precyzyjnie zmierzyli przebieg prostego, spontanicznego aktu wolicjonalnego, byli dwaj niemieccy neurolodzy Hans Kornhuber i Lüder Deecke. Podczas eksperymentu polegającego na mierzeniu czasu reakcji mózgu na nagłe zgięcia nadgarstka zauważyli oni, że – zanim nastąpi jakikolwiek ruch, to z dużym wyprzedzeniem pojawia się związana z nim reakcja mózgu. Hans Kornhuber i Lüder nazwali tę podwyższoną aktywność potencjałem gotowości (*Bereitschaftspotential, readiness potential* – RP) (Kornhuber & Deecke, 1965). Udało się – na skutek zastosowania elektromiografu do detekcji ruchu mięśni nadgarstka oraz elektroencefalografu do pomiaru aktywności elektrycznej mózgu – precyzyjnie zmierzyć czas pomiędzy początkiem RP a pierwszym skurczem mięśni. Statystycznie RP wyprzedzało skurcz mięśni o ok. 0,8 sekundy. Wykres obrazujący opisane powyżej zjawisko wygląda następująco:



Rys. 2. Przebieg potencjału gotowości: zmiana pola elektrycznego określonych obszarów mózgu następuje ok. 0,8 sekundy przed wykonaniem czynności.

Potencjał gotowości osiąga maksimum ok. 90 milisekund przed wykonaniem czynności. Analiza aktywności półkul mózgowych pokazała z kolei, że ok. 50 milisekund przed akcją motoryczną aktywowany jest obszar mózgu odpowiedzialny za „wysterowanie" mięśnia – aktywację tę zwykło się nazywać potencjałem ruchu (*movement potential*).

Powyższe wyniki odbiły się szerokim echem w środowisku uczonych badających działanie mózgu oraz świadomość. Po raz pierwszy udało się prześledzić czasowy przebieg prostego aktu wolicjonalnego. Eksperyment ten nie zmienił, mimo zaskakującego efektu (aktywacja mózgu reprezentująca gotowość do ruchu dokonuje się znacznie wcześniej, niż inicjacja samej czynności), sposobu myślenia o funkcjonowaniu samego aktu wolicjonalnego. Relacja między podmiotem i jego wolą a realizowaną czynnością była zachowana w sensie następstwa czasowego, tzn. najpierw pojawiało się zdarzenie mentalne (zamiar wykonania czynności), a dopiero później czynność. Neuropsychologiczne badania Benjamina Libeta wyznaczają kolejny etap w rekonstrukcji przebiegu prostych działań intencjonalnych.

Założenia metodologiczne wybranych eksperymentów Libeta

Benjamin Libet jest autorem kilku znaczących eksperymentów. W kontekście niniejszej pracy zaprezentowane zostaną tylko te, które wiążą się z ustaleniem czasowego przebiegu aktu wolicjonalnego. Libet, przystępując do badań w latach 50-tych XX wieku, przyjął, iż z zasady korelatów psychoneuronowych wynikają dwa postulaty epistemologiczne, które należy potraktować jako ramowe założenia podczas przeprowadzania eksperymentów neurofizjologicznych:

- po pierwsze, należy uznać, że raporty introspekcyjne są (tzw. samoopisy (*self-report*)) niezbędnym kryterium operacyjnym;
- po drugie, nie należy *a priori* zakładać określonego charakteru relacji: mózg – umysł, zwłaszcza takiego, który eliminowałby znaczenie poziomu mentalnego (Libet, 2004, s. 18).

Libet przyjął dodatkowo, że wola ma charakter endogenny, czyli powstaje pod wpływem przyczyn wewnętrznych, nie zaś zewnętrznych – i w tym sensie badanie musi być tak skonstruowane, by czynniki zewnętrzne lub patologiczne nie wpływały na przebieg eksperymentów. Wśród schorzeń eliminujących wiarygodne badanie woli wymienił on m.in. zespół Touretta, który wywołuje działania mimowolne.

Półsekundowe opóźnienie

Pierwszym eksperymentem Libeta, który wywołał szeroką dyskusję w kręgach neurofizjologicznych, był pomiar długości trwania stymulacji elektrycznej w obszarze

czuciowej kory somatosensorycznej odpowiedzialnej za powstanie świadomego wrażenia (samo badanie odbywało się w czasie zabiegu operacyjnego z miejscowym znieczuleniem). Zgodnie z wynikami badań zgromadzonymi do początku lat 80. XX wieku wiedziano, że sygnał czuciowy przekazywany jest ze skóry – poprzez rdzeń kręgowy – kilkoma szlakami nerwowymi do mózgu. Niestety, ówczesny poziom narzędzi diagnostycznych uniemożliwiał precyzyjne zarejestrowanie czasu, w którym sygnał docierałby do mózgu. Przyjęta przez Libeta strategia bezpośredniego pobudzania kory czuciowej pozwalała ominąć to ograniczenie.

Stymulacje wykonywane w trakcie eksperymentu były opisywane przez cztery zmienne: natężenie prądu, długość trwania pojedynczego impulsu, częstotliwość impulsów oraz czas trwania pobudzenia. W serii pomiarów postanowiono określić, jak poszczególne zmienne wpływają na czas, w którym pojawi się świadome wrażenie. Po statystycznym opracowaniu wyników Libet przedstawił następujące wnioski:

1. potrzebny jest ciąg impulsów, aby wzbudzić słabe (progowe) wrażenie, trwający przynajmniej ok. 0,5 sekundy; pojedynczy, nawet silny impuls nie potrafi wywołać świadomego wrażenia; podobny efekt uzyskano pobudzając nerwy wzgórza (*thalamus*) – tu również okazało się, że potrzebny jest ciąg impulsów trwający co najmniej 0,5 sekundy, by wywołać świadome wrażenie[49];
2. istnieje graniczna częstotliwość impulsów elektrycznych, poniżej której nie powstają świadome wrażenia; zwiększenie tej częstotliwości powoduje, że możliwe jest obniżenie amplitudy natężenia impulsów, nie wpływa to jednak na czas trwania ciągu progowego (niczego już nie zmienia zwiększanie częstotliwości impulsów przy minimalnym natężeniu);
3. istnieje graniczne natężenie prądu, poniżej którego nie jest możliwe wywołanie świadomego wrażenia, z kolei wzrost tego natężenia (w dopuszczalnych granicach) powoduje nasilenie wrażeń oraz skrócenie się czasu ich powstawania (zbyt intensywne natężenie może powodować wzbudzenie większej liczby neuronów i

[49] Wynik ten potwierdziło później kilka grup badawczych, przy czym jedna wykazała, że możliwe jest skrócenie ciągu impulsów potrzebnych do wywołania świadomego wrażenia do 0,25 sekundy. Tak wyraźne skrócenie czasu mogło mieć związek, zdaniem Libeta, z pomiarem przeprowadzonym na pacjentach chorych na epilepsję. P. G. Ray, K. J. Meador, C. M. Epstein, D. W. Loring, L. J. Day, *Magnetic stimulation of visual cortex: Factors influencing the perception of phosphenes*, [w:] „Journal of Clinical Neurophysiology", 15(4), str. 351–357. Cyt. za: B. Libet, *Time Factors in Conscious Processes: Reply to Gilberto Gomes*, „Consciousness and Cognition" 9, 1–12 (2000).

prowadzić do wywołania sztucznej – z punktu widzenia życia codziennego – sytuacji).

Pierwszy wniosek jest najbardziej intrygujący w kontekście niniejszych rozważań. Wskazuje on na stosunkowo długi czas aktywowania struktur odpowiedzialnych za powstanie świadomego wrażenia. W związku z tym nasuwa się następujące pytanie: skoro minimalny czas potrzebny do wywołania zdarzenia mentalnego wynosi 0,5 sekundy, to czy każde tego typu wrażenie jest opóźnione o tę jednostkę czasu w stosunku do czasu rzeczywistego (umysł niejako antydatuje zdarzenia w porównaniu z faktycznym momentem ich rozpoczęcia)? Odpowiedź Libeta jest w tym kontekście twierdząca, jednak szczegółowe wyjaśnienie tej kwestii wymagałoby przytoczenia kolejnych danych eksperymentalnych (nieistotnych w perspektywie niniejszej pracy). Dla funkcjonowania działań intencjonalnych istotne jest tylko to, że każde świadome wrażenie wymaga, zdaniem Libeta, pobudzenia odpowiednich struktur w mózgu, które trwa co najmniej 0,5 sekundy.

Na potwierdzenie powyższego wyniku amerykański psycholog przytacza badania Arthura Jensena (1979) dotyczące czasu reakcji. W eksperymencie tym proszono uczestników o jak najszybsze naciśnięcie przycisku w reakcji na usłyszany sygnał dźwiękowy. Uzyskane czasy reakcji należały do przedziału od 200 do 300 ms. Ponieważ Jansen miał wątpliwość, czy wielkość tego przedziału nie jest następstwem faktu, iż niektórzy badani wykorzystują w trakcie eksperymentu myślenie refleksyjne, poprosił ich, by powtórzyli badania, ale w taki sposób, aby minimalnie pogorszyć uzyskany w pierwszej serii wynik. Efekt był taki, że przedział w zaskakujący sposób zwiększył się – tym razem najkrótsze czasy reakcji wynosiły 600 ms, a najdłuższe 800 ms. Rezultat ten, zdaniem Libeta, był skutkiem faktu, iż chęć minimalnego pogorszenia czasów reakcji angażowała procesy świadomej kontroli, co w efekcie musiało wywołać przesunięcie o ok. 500 ms.

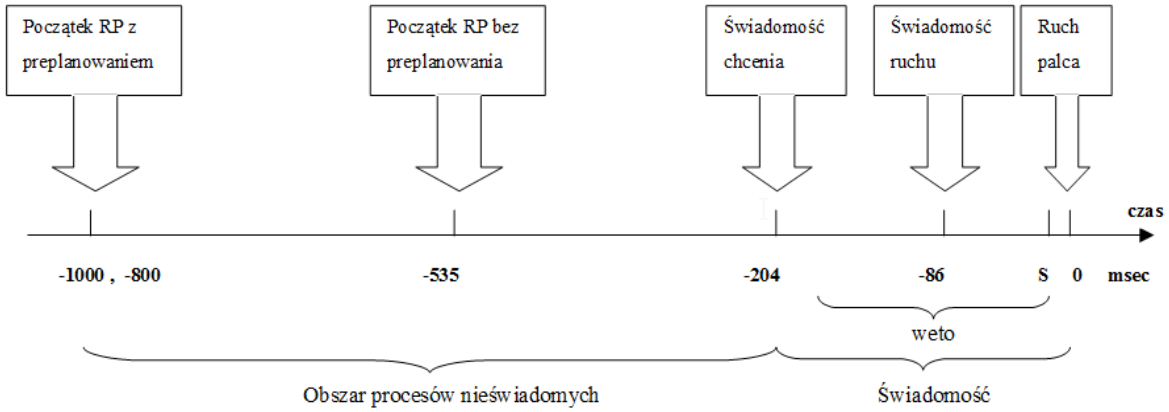
0,5 sekundy opóźnienie w kontekście działania intencjonalnego

Kornhuber i Deecke wykazali, że potencjał gotowości poprzedza skurcz mięśni w trakcie nagłego ruchu nadgarstka o ok. 0,8 s. Przy zestawieniu tego wyniku z rezultatem Libeta wyłania się następujące pytanie: czy w ramach działania intencjonalnego, towarzysząca mu chęć zainicjowania ruchu pojawia się zaraz na początku, czy może, zgodnie z zasadą półsekundowego opóźnienia, powstaje dopiero na pewnym jego etapie?

Aby odpowiedzieć na powyższe pytanie, potrzebny jest sposób, za pomocą którego można wskazać moment pojawienia się chęci wykonania działania. Metoda polegająca na poinformowaniu eksperymentatora o zaistnieniu tego typu chęci lub przyciśnięcie przycisku z pewnością zaburzałyby uzyskany wynik, gdyż prowadziłyby do inicjowania dodatkowych działań. Rozwiązanie zaproponowane przez Libeta polegało na wprowadzeniu do eksperymentu specjalnie spreparowanego zegara. Funkcję tę pełnił oscyloskop z plamką o okresie obiegu po okręgu symulującym tarczę zegara wynoszącym 2,56 sek. Równocześnie rozszerzono cel zadania: osoba biorąca udział w eksperymencie, poza ruchem palca, miała również zapamiętać położenie kropki na zegarze Wundta w chwili, gdy uświadomi sobie chęć wykonania ruchu. Zniesiono także ograniczenia, które narzucili w swoim eksperymencie Kornhuber i Deecke, tj. limit czasowy na wykonanie zadania (w tym ostatnim przypadku instrukcja polecała wykonanie kilku ruchów nadgarstka w ciągu określonego czasu). Libet zrezygnował z tego ograniczenia, gdyż, jak sądził, wprowadzało ono do eksperymentu dodatkowy efekt, tzw. preplanowania, polegającego na wykonywaniu zadania we wcześniej zaplanowanych momentach, np. zawsze, gdy wskazówka zegara znajdzie się na godz. 12-tej.

Wyniki

Wyniki, które Libet uzyskał w 1983 roku, można zobrazować w następujący sposób (Libet i in., 1983b):



Rys. 3. Przebieg aktu wolicjonalnego dla ruchu palca na podstawie eksperymentu Libeta. Wskazane na osi czasu wartości reprezentują zdarzenia charakterystyczne dla prostego działania intencjonalnego. Są to odpowiednio: (1) początek narastania potencjału gotowości w przypadku działania planowanego (np. poruszę palcem, gdy wskazówka zegara pokaże godzinę 12:00), (2) początek narastania potencjału gotowości dla działań w pełni spontanicznych, (3) moment uświadomienia sobie chęci wykonania ruchu, (4) moment uświadomienia sobie, że właśnie zaczynam wykonywać ruch, (5) S to zdarzenie odnoszące się do zewnętrznego impulsu sensorycznego, przy pomocy którego Libet kontrolował 'stronniczość' (bias) uczestnika w odczytywaniu pozycji zegara

(jeśli moment odczytu S znacząco różnił się od faktycznego zdarzenia sensorycznego, uczestnik był o tym informowany i proszony o zwiększenie uwagi podczas realizacji eksperymentu), (6) moment, gdy elektromiograf zarejestrował początek procesu odpowiedzialnego za realizację ruchu.

Powyższy rysunek prezentuje uśredniony przebieg prostego działania intencjonalnego od momentu, gdy zaczyna narastać potencjał gotowości (RP), do chwili wykonania ruchu dla trzech niezależnych serii (każda po 40 prób). Pierwsza seria (W) dotyczyła pomiaru, w którym badany miał zarejestrować na zegarze moment uświadomienia sobie chęci wykonania ruchu. W drugiej serii (M) badany miał zarejestrować chwilę, w której uświadomi sobie, że właśnie poruszył palcem (*actually moved*). W ostatniej serii (S) badany miał zarejestrować moment, w którym doświadczył lekkiego pobudzenia sensorycznego wywołanego stymulacją po zewnętrznej stronie dłoni. Rezultaty uzyskane przez Libet można podsumować w następujący sposób:

> *Wyniki badań sugerują, że mózg jako pierwszy zaczyna coś robić (nie wiemy tylko, co to jest). Następnie osoba staje się świadoma chęci wykonania działania. To byłoby miejsce, w którym ujawnia się świadoma wola, przynajmniej w takim znaczeniu, że osoba po raz pierwszy staje się świadoma, że próbuje coś zrobić. Następnie, tuż przed wykonaniem ruchu – osoba zdaje sobie sprawę z tego, że palec się porusza. Wreszcie palec wykonuje ruch.*[50]

Warto zwrócić uwagę na znaczenie preplanowania w przebiegu działania. Widać wyraźnie, że jeśli badany planuje jakąś czynność (np. poruszenie palcem, gdy wskazówka znajdzie się na godzinie 12.), to potencjał gotowości zaczyna narastać znacznie wcześniej, niż wtedy, gdy ruch jest również swobodny, ale decyzja o jego wykonaniu jest całkowicie spontaniczna.

Wnioski z eksperymentu Libeta[51]

[50] "These findings suggest that the brain starts doing something first (we don't know just what that is). Then the person becomes conscious of wanting to do the action. This would be where the conscious will kicks in, at least, in the sense that the person first becomes conscious of trying to act. Then, and still a bit prior to the movement, the person reports becoming aware of the finger actually moving. Finally, the finger moves." (Wegner, 2002, s. 53).

[51] Dla pełnego obrazu należy odnotować, że poza grupą badaczy, którzy potwierdzili wyniki Libeta (Keller i Heckhausen w 1990 oraz Haggard i Eimer w 1999), są również tacy badacze , którzy kwestionują poprawność uzyskanych w paradygmacie Libeta wyników (Gomes, 1998; Pockett & Miller, 2007) lub ich interpretację (Mele, 2009). Główny zarzut odnosi się do sposobu określania momentu, w którym pojawia się świadomość chcenia oraz sposobu funkcjonowania mechanizmu weta. Haggard zwraca np. uwagę na to, iż Libet nie uwzględnił w swoim eksperymencie opóźnienia, które powstaje, kiedy badany musi dzielić uwagę pomiędzy obserwacją palca a obserwacją zegara. W jego opinii takie dzielenie czy wręcz przełączanie uwagi

Najbardziej zaskakujące i przełomowe – w odniesieniu do tradycyjnego podejścia – jest ustalenie momentu czasowego reprezentującego „uświadomienie sobie chęci wykonania ruchu". Moment ten w przypadku ruchu bez preplanowania pojawia się ok. 330 ms po pojawieniu się potencjału gotowości, a ok. 200 ms przed samym ruchem. Takie usytuowanie w czasie wskazuje, iż w przebiegu tego typu aktów znaczącą fazę stanowią procesy nieświadome reprezentowane przez narastające RP, natomiast świadoma chęć zainicjowania działania pojawia się dopiero na późniejszym etapie. Wniosek ten, zdaniem Libeta, potwierdzają wszystkie sytuacje, w których pojawia się tzw. automatyzacja czynności, czyli taki tryb realizacji zachowań, w którym określone stany świadomości pełnią funkcję kontrolną w ograniczonym zakresie.

Drugim, zaprezentowanym na powyższym wykresie, istotnym elementem jest okres oznaczony słowem „weto". Rozciąga się on od chwili powstania świadomości chcenia do chwili 'M' znajdującej się średnio 86 milisekund przed wykonaniem ruchu (Libet i in., 1983b). Libet przypisuje temu okresowi duże znaczenie. Jest to czas, w którym, na skutek oddziaływania naszej świadomości, możemy efektywnie wycofać się z „zaplanowanej" przez procesy nieświadome czynności. Weto to inaczej świadomy i nieuwarunkowany kontroler działań przygotowywanych przez procesy nieświadome. Danymi wejściowymi dla mechanizmu weta są nie tylko rezultaty procesów związanych z aktem wolicjonalnym, ale również efekty innych procesów mentalnych realizowanych w tym samym czasie (np. spostrzeganie). Przykładem zastosowania weta może być następująca scena: pod wpływem złych emocji chcemy wypowiedzieć coś niecenzuralnego pod adresem osoby X. Nagle, gdy prawie zaczynamy mówić, spostrzegamy, że osoba X przechodzi obok nas – w tym momencie, zdaniem Libeta, mamy jeszcze możliwość zawetowania swojej wypowiedzi i powstrzymania się od jej wygłoszenia. Można jednak spytać: skoro działania intencjonalne w dużym stopniu inicjowane są przez procesy znajdujące się poza świadomą kontrolą, to dlaczego mechanizm weta również nie podlega tego typu kontroli? Wyobrażalne jest bowiem takie ujęcie, w którym weto również kontrolowane jest przez procesy nieświadome. Libet zdecydowanie odrzucał taką możliwość. Twierdził on, że uzyskane

może pochłonąć określony czas który należałoby uwzględnić określając położenie intencji w relacji do RP (Haggard i in., 2002). Zarzuty Pockett i Millera dotyczą opóźnienia związanego z uświadomieniem sobie położenia punktu na zegarze. Na wszystkie te zarzuty Libet starał się odpowiedzieć, kwestionując ich zasadność lub przywołując wyniki innych badaczy, którzy stosowali podobny paradygmat eksperymentalny i wielokrotnie potwierdzili, że uzyskany przez niego wynik dotyczący relacji „RP – moment uświadomienia sobie chęci wykonania ruchu" jest prawidłowy. Ostatecznie jednak kluczowe jest to, że sugerowane błędy w żadnym przypadku nie były tak poważne, by zakwestionować ogólny schemat wskazujący, że procesy nieświadome poprzedzają pojawienie się świadomości (Gomes, 1998).

przez niego dane eksperymentalne tego nie potwierdzają. Zaproponowana przez Libeta interpretacja mechanizmu weta jest eksperymentalnie słabo potwierdzona. On sam przyznaje, że to, co udało się zbadać, to możliwość zawetowania czynności wcześniej zaplanowanej, natomiast w przypadku ruchów spontanicznych nie udało się przeprowadzić eksperymentów potwierdzających tego typu ujęcie. Z drugiej strony, istnieją eksperymenty wykonane np. przez Gordona Logana (Logan, 1994) lub Hakwana C. Lau wraz ze współpracownikami (Lau i in., 2007), które wskazują na duże ograniczenia mechanizmu weta. Na ich podstawie można nawet zaryzykować hipotezę, że weto jest tylko szczególnym przypadkiem działania mechanizmu odpowiedzialnego za korygowanie zainicjowanych działań. Z badań Yvesa Rossetti'ego (Pisella i in., 2000), dotyczących możliwości świadomego powstrzymywania ruchu ręki, wynika, że kiedy taki ruch jest realizowany, to nie można go zatrzymać w dowolnie wybranym momencie. W tego typu przypadkach człowiek dysponuje jedynie możliwością ograniczonej korekty. Nie jest to eksperyment rozstrzygający, gdyż odnosi się nie do czynności, która miałaby zostać zainicjowana (to jej dotyczy Libetowskie weto), lecz do już trwającej. Wynik Rosettiego wskazuje jednak na to, że powstrzymanie się od działania (zawetowanie go) jest możliwe tylko po spełnieniu ściśle określonych warunków.

Intencja w działaniu jako korelat procesów przygotowawczych

Przedstawione powyżej wnioski nie zostały w pełni zaakceptowane przez naukową społeczność. Mechanizm weta traktuje się obecnie bardziej jako historyczną ciekawostkę, niż wymagającą dalszych badań, intrygującą hipotezę. Problem relacji między mózgowymi procesami przygotowującymi do działania (reprezentowanymi przez narastający potencjał gotowości) a świadomie doświadczaną intencją zainicjowania ruchu spowodował żywą dyskusję w środowisku neuronaukowców i filozofów (D. Dennett & Kinsbourne, 1992; Mele, 2009) i przyczynił się do powstania szeregu projektów badawczych, których celem było doprecyzowanie wskazanej relacji (Gomes, 1998; Haggard, 2008; Pockett i in., 2006).

Uzyskany przez Libeta wynik stał się od chwili jego ogłoszenia jednym z najczęściej przytaczanych empirycznych argumentów przywoływanych przeciwko istnieniu wolnej woli. Skoro bowiem świadoma intencja pojawia się z tak znaczącym opóźnieniem w porównaniu z rejestrowalnymi empirycznie mózgowymi procesami przygotowującymi organizm do wykonania ruchu, to trudno uznać, że to ona właśnie jest przyczyną działania (Pockett i in., 2006). Wielu filozofów nie zgadza się z tego typu argumentacją i próbuje,

odmiennie interpretując dane eksperymentalne, „odzyskać" efektywność intencji (Mele, 2009, s. vii). Choć sam spór między zwolennikami a przeciwnikami wolnej woli jest interesujący, to z perspektywy niniejszej pracy ważniejsze okazuje się rozważenie alternatywy, sformułowanej przez Patricka Haggarda.

> *Nowatorskie badania Benjamina Libeta sugerowały, że świadoma intencja pojawia się po wystąpieniu przygotowawczej aktywności mózgu. W związku z tym nie może ona być przyczyną naszych działań, ponieważ przyczyna nie może występować po swoim skutku. W tej sytuacji pozostają dwie możliwości. Albo świadoma intencja mogłaby być częścią iluzyjnego wyobrażenia przyczynowości mentalnej, wywnioskowanego post factum, aby wyjaśnić zachowanie, albo świadoma wola byłaby bezpośrednim skutkiem procesów mózgowych, które przygotowują działanie. Zgodnie z tym [ostatnim] ujęciem, intencja jest świadomościowym korelatem neuronalnej aktywności procesów przygotowawczych.*[52]

Wskazana przez Haggarda alternatywa została poddana weryfikacji empirycznej. Obecnie uważa się, że intencja w działaniu (rozumiana tak, jak definiuje się ją w instrukcji do eksperymentu Libeta) ma status świadomościowego korelatu procesów przygotowujących ruch (np. ruch palca, nadgarstka lub ręki), nie można zatem mówić o relacji przyczynowej pomiędzy RP a intencją. Do takiego wniosku skłaniają, zdaniem Patricka Haggarda, dane z dwóch eksperymentów. W pierwszym z nich – zrealizowanym zgodnie z instrukcją Benjamina Libeta przy wykorzystaniu skanera fMRI – udało się zidentyfikować struktury zaangażowane w powstanie intencji ruchu. Korelat świadomościowy intencji wyznaczony został poprzez odjęcie aktywacji mózgu zidentyfikowanej dla Libetowskiego „warunku W"[53] od aktywacji dla „warunku M". W ten sposób udało się określić, że w powstanie tego typu intencji zaangażowane są następujące struktury: przednia część dodatkowego pola ruchowego (*pre-supplementary motor area* – pre-SMA) oraz bruzda śródciemieniowa (*intra-parietal sulcus*) (Lau, Rogers, Haggard, & Passingham, 2004). W drugim

[52] "The seminal studies of Benjamin Libet suggested that conscious intention occurs after the onset of preparatory brain activity. It cannot therefore cause our actions, as a cause cannot occur after its effect. Two other possibilities remain. Either conscious intention could be part of an illusion of mental causation, retrospectively inferred to explain behavior. Alternatively, conscious intention could be an immediate consequence of the brain processes which prepare action. On this view, intention is a conscious correlate of preparatory neural activity." (Haggard, 2005, s. 291).

[53] Warunki W (*willing*) oraz M (*movement*) odnoszą się do różnych etapów realizacji działania. Pierwszy dotyczy momentu, w którym badany uświadamia sobie chęć wykonania ruchu, a drugi chwili, w której rozpoczyna się realizacja ruchu. Zgodnie z eksperymentem Libeta oraz jego replikacjami pomiędzy W a M występuje odstęp czasowy równy ok. 250 ms.

eksperymencie, również opartym na instrukcji Libeta, tym razem z użyciem EEG i EMG, Angela Sirigu – wraz ze współpracownikami – zaobserwowała, że osoby z lezją w okolicach bruzdy śródciemieniowej wykazują znaczące opóźnienie czasowe w ocenie „warunku W" w porównaniu z grupą kontrolną. Standardowo intencja wyprzedza ruch o ok. 200 ms, natomiast w przypadku osób z lezją okres ten skraca się do zaledwie 55 ms (Sirigu i in., 2003).

Wyniki wymienionych eksperymentów, zdaniem Haggarda, potwierdzają następujący pogląd:

> *[...] płaty czołowy oraz ciemieniowy wspólnie tworzą obwód, którego zadaniem jest opracowywanie i monitorowanie planów motorycznych przyszłych zachowań – intencja w tym obwodzie to jeden z elementów realizowanej symulacji*. (Haggard, 2005, s. 292).

W tym kontekście nasuwają się pytania: Jaką funkcję pełni intencja w działaniu? Do czego służy tego typu świadomościowy korelat?

Autorzy drugiego z wymienionych eksperymentów wysunęli ciekawą hipotezę dotyczącą tego zagadnienia. W ich opinii:

> *Aby chciane działanie było zachowaniem funkcjonalnym, mózg musi być wyposażony w mechanizm umożliwiający dopasowanie skutków działania ruchowego do prior intencji*.[54] (Sirigu i in., 2003).

Zaproponowane przez nich podejście sugeruje interesującą możliwość, otóż system decydujący o realizacji zachowań celowych posługuje się dwoma rodzajami reprezentacji. Z jednej strony są to określone stany intencjonalne (np. prior intencje, przekonania, pragnienia), a z drugiej reprezentacje w formie planów motorycznych, które – choć są zależne od wskazanych stanów – nie są ich prostym odwzorowaniem. W tym układzie, intencja w działaniu stanowiłaby, zgodnie z sugestią badaczy, rodzaj świadomościowej reprezentacji wybranych części mózgowych planów motorycznych i w ten sposób zapewniałaby agentowi możliwość kontroli dopasowania między zamiarem a realizowanymi przez procesy mózgowe zachowaniami (bardziej szczegółowe omówienie tego elementu znajdzie się w ostatnim rozdziale pracy).

[54] *For willed action to be a functional behavior, the brain must have a mechanism for matching the consequences of the motor act against the prior intention* (Sirigu i in., 2003).

Podwójna treść intencji w działaniu

Innym ważnym rezultatem badań psychologii intencji jest zidentyfikowanie dwóch treściowych składowych intencji w działaniu. Warto przypomnieć, że John Searle wyróżnił w swoim aparacie pojęciowym intencje proste oraz intencje złożone. Pierwsze dotyczą tzw. działań prostych (np. chcę pociągnąć za spust w pistolecie, drugie natomiast odnoszą się do działań złożonych, nadbudowanych niejako nad działaniami prostymi. Charakteryzują się one wielopoziomowym opisem oraz tzw. efektem akordeonu, tj. treść intencji można rozłożyć na szereg powiązanych ze sobą składowych, uporządkowanych ze względu na relacje przyczynowe (‘za pomocą’) lub relacje konstytuowania (‘poprzez’).

O ile ciągle skazani jesteśmy na mniej lub bardziej subtelne analizy pojęciowe w przypadku działań złożonych, o tyle w przypadku działań prostych dysponujemy danymi eksperymentalnymi pomocnymi w odsłonięciu struktury intencji. Oprócz wyników eksperymentów opartych na instrukcji Libeta, możemy sięgnąć także po raporty osób poddanych elektrycznej stymulacji mózgu oraz po wyniki badań dotyczących zachowań naśladowczych. Na ich podstawie można skonstruować następujący obraz intencji: treść intencji odnoszącej się do działań prostych (pojedynczych ruchów lub zautomatyzowanych umiejętności) zawiera dwa komponenty: (1) poczucie chęć wykonania ruchu (*sense of urge or being about to move*) oraz (2) odniesienie do docelowego obiektu lub zdarzenia (*reference forward to the goal object or event*) (Haggard, 2005). Poniżej omówię każdą z tych składowych z osobna.

Poczucie chęci wykonania ruchu – pierwsza składowa treściowa intencji w działaniu

Pierwszy komponent ma charakter egocentryczny, tzn. skierowany jest na własne ciało i własne odczucia, zawiera też zgrubną reprezentację planowanego ruchu cielesnego. Dobrym przykładem, w którym tego typu składowa jest widoczna, okazuje się uświadomiony stan gotowości do natychmiastowego wykonania ruchu. Pojawienia się chęci do ruchu „wypatrywać” mają uczestnicy eksperymentu zgodnego z instrukcją Libeta, a dokładnie z tzw. warunkiem W. Innym przykładem potwierdzającym istnienie tego typu składowej są – w związku z zabiegiem neurochirurgicznym, który ma ograniczyć niekontrolowane napady epilepsji – raporty osób poddanych bezpośredniej stymulacji elektrycznej w okolicach dodatkowego obszaru przedruchowego (SMA) (Fried i in., 1991). Fried wraz z zespołem współpracowników – poprzez implantację macierzy elektrod w

okolice SMA – przeprowadził systematyczne mapowanie określonych miejsc kory oraz odwzorował reakcje motoryczne przy wykorzystaniu stymulacji elektrycznej. Fried zgromadził również, oprócz mapowania: „stymulacja – ruch kończyny", towarzyszące stymulacjom subiektywne raporty pacjentów. Poszczególne relacje podzielone zostały na trzy grupy:

1. wrażenia odbierane jako uczucie mrowienia, drętwienia, ciepła oraz lekkiego bólu,
2. subiektywne wrażenie, że badany wykonał ruch przy równoczesnej nieobecności jakiejkolwiek aktywności ruchowej,
3. subiektywne poczucie chęci (*urge*) wykonania ruchu lub jego antycypacji.

Ostatni ze wskazanych przypadków wprost odnosi się do motorycznej, zgrubnej reprezentacji ruchu należącej do składowej intencji w działaniu. Fried odnotował także, co szczególnie warto podkreślić, że w pewnych sytuacjach zwiększona stymulacja (u osób odczuwających chęć poruszenia kończyną) prowadziła do jawnej odpowiedzi (reakcji) motorycznej, przy czym nie zawsze była to odpowiedź zgodna z wywołaną przez stymulację treścią intencji (zdarzało się, że zwiększone natężenie prądu prowadziło do ruchu innej części ciała niż ta, która dostępna była w treści intencji) (Fried i in., 1991).

Odniesienie do docelowego obiektu lub zdarzenia – druga składowa intencji w działaniu

Intencja, oprócz chęci wykonania ruchu, zawiera treść odnoszącą się do zewnętrznego celu, ze względu na który dane działanie zostało podjęte. William James wyraźnie wyeksponował ten element w koncepcji działań ideomotorycznych i dobrowolnych. W jego opinii, na poziomie idei (współcześnie powiedzielibyśmy: „reprezentacji"), ważniejszą funkcję pełnią efekty działań, niż konkretne zachowania, które prowadzą do ich osiągnięcia (Haggard, 2005, s. 293). Ta część intencji w działaniu, która pełni funkcję odniesienia do celu, jest w tym kontekście jednym z kluczowych mechanizmów kontroli zachowań.

Potwierdzeniem tego typu tezy są wnioski płynące z badań dotyczących zachowań naśladowczych u dzieci. Okazuje się, że jeśli poprosi się dzieci w wieku od 3 do 6 lat o powtórzenie pewnej prostej sekwencji ruchów (np. złapanie lewą ręką prawego ucha), to można zaobserwować intrygujący wzorzec: pewne sekwencje są realizowane ze znacząco większą liczbą błędów niż inne (Bekkering i in., 2000). Mniej błędów jest popełnianych, gdy ucho i ręka znajdują się po tej samej stronie, np. dziecko ma chwycić lewą ręką lewe

ucho. Efekt ten badacze z Max-Planck Institute for Psychological Research w Monachium wyjaśniają w następujący sposób: w momencie, gdy dziecko obserwuje sekwencję ruchów dorosłego równocześnie próbuje zidentyfikować leżący u jej podstaw cel, który następnie stara się osiągnąć podczas powtórzenia. Przy takim założeniu, twierdzą autorzy eksperymentu, mniej istotne są dla dziecka konkretne ruchy (np. podniesienie prawej lub lewej ręki), a bardziej liczy się końcowy efekt (np. chwycenie lewego ucha). Innymi słowy, naśladownictwo nie jest prostym odtworzeniem zaobserwowanych ruchów. W tego typu zadaniach dziecko może się skupiać na celu i modyfikować naśladowane zachowanie, aby łatwiej osiągnąć pożądany skutek.

Podobny efekt wykazały badania dotyczące ruchów gałek ocznych, które zarejestrowano podczas realizacji lub obserwacji działania polegającego na przenoszeniu drewnianych klocków z miejsca A do B (Flanagan & Johansson, 2003). Po przeanalizowaniu zebranych danych okazało się, że wzrok uczestników zarówno podczas obserwacji, jak i wykonania działania wyraźnie realizował wzorzec predykcyjny, tzn. niemal całkowicie koncentrował się na przewidywanych miejscach kontaktu dłoni z klockiem oraz ich docelowej pozycji. Tego typu dane świadczą o tym, zdaniem Flanagana, że podczas obserwacji działania ruch gałek ocznych realizuje program sterowany przez antycypowany skutek zawarty w treści motorycznej reprezentacji. W związku z tym efektem zaproponowano hipotezę bezpośredniego dopasowania (*the direct matching hypothesis*), zgodnie z którą rozumienie działań manualnych odbywa się poprzez ich łączenie z odpowiadającymi im reprezentacjami motorycznymi. Innymi słowy, gdy patrzę na kogoś, kto przygotowuje herbatę, to niejako automatycznie zaczynam skupiać wzrok na elementach, które w przyszłości pozwolą mi wykonać tego typu działanie. Z przedstawionych danych i analiz można wyciągnąć jeszcze jeden wniosek: podczas realizacji działania lub jego obserwacji kluczowe są dla nas efekty poszczególnych ruchów (ich sensoryczny wymiar), a mniej ich motoryczna podstawa. Jest to zatem kolejne potwierdzenie złożonej natury intencji w działaniu, która zawiera, oprócz poczucia chęci wykonania określonego ruchu (*sense of urge*), składową celowościową, która w dużej mierze wpływa na kształt danego działania (*reference forward to the goal object or event*).

W dotychczasowych rozważaniach skupiłem się głównie na identyfikacji składowych intencji, nie odnosiłem się natomiast do relacji między nimi. Można powiedzieć, odnosząc przedstawione analizy do Searle'a teorii intencjonalności, że warunki spełniania składowej motorycznej oraz składowej celowościowej w dużym stopniu są niezależne. Poczucie chęci

wykonania ruchu odnosi się głównie do określonego stanu ciała agenta, natomiast odniesienie do docelowego obiektu lub zdarzenia posiada głównie warunki spełniania dotyczące oczekiwanego stanu świata. Z drugiej strony, składowa motoryczna powinna w jakiś sposób określać składową efektywnościową, choćby dlatego, że planowane ruchy powinny, przynajmniej w pewnym zakresie, doprowadzić do realizacji oczekiwanych rezultatów. Potwierdzenie, że tego typu związek istnieje, znaleźć można w badaniach Patricka Haggarda dotyczących percepcji obu składowych. Badacz ten postanowił porównać w dwóch sytuacjach sposoby, za pomocą których odbieramy moment pojawienia się zdarzenia sensorycznego (sygnału dźwiękowego): kiedy jest on powiązany z działaniem oraz kiedy jest niezależny od działania. Okazało się, że jeśli tego typu zdarzenie nie jest w żaden sposób powiązane z działaniem intencjonalnym, wówczas postrzegamy je jako późniejsze w porównaniu z tym samym zdarzeniem, ale wygenerowanym jako skutek działania. Innymi słowy, jeśli działanie prowadzi do obserwowalnych zmian w środowisku, to postrzegamy je inaczej pod względem czasowym, niż to samo działanie, ale pozbawione wpływu na otoczenie. Jest to tzw. efekt scalania (*binding*), obejmujący prawdopodobnie wszystkie zachowania percypowane przez sprawcę jako związki przyczynowo-skutkowe (Patrick Haggard i in., 2002).

Haggard postanowił sprawdzić, dysponując tego typu efektem, czy występuje on w sytuacji, gdy działanie nie jest w pełni dowolne, tak by móc zweryfikować wpływ intencji na efekt wiązania. W projektowanym układzie eksperymentalnym zastosowano urządzenie do przezczaszkowej stymulacji magnetycznej (TMS). Badani, podobnie jak w eksperymencie mierzącym efekt scalania, mieli za zadanie określić moment, w którym w wybranych przez siebie chwilach nacisnęli przycisk lub usłyszeli dźwięk wygenerowany w związku z jego naciśnięciem. Jedyna różnica polegała na tym, że w losowo wybranych próbach, kiedy badany przygotowywał się do poruszenia palcem, wytwarzany był impuls magnetyczny, który powodował mimowolny ruch palca i w konsekwencji – naciśnięcie klawisza. W obu przypadkach badany percypował związek przyczynowo-skutkowy (naciśnięcie klawisza → pojawienie się dźwięku), ale tylko w jednym z nich „aktywna" była intencja. Okazało się, po przeanalizowaniu danych dla prób bez „aktywnej" intencji, że nie występuje efekt scalania (Patrick Haggard i in., 2002). Uzyskany rezultat, zdaniem Haggarda, skłania do wyciągnięcia dwóch wniosków. Po pierwsze, intencja w działaniu to bardzo istotny element układu odpowiedzialnego za realizację działań dobrowolnych. Jej brak prowadzi do istotnych zmian w percepcji przebiegu działania. Po drugie, intencja, w

odróżnieniu od poczucia sprawstwa (patrz: analiza poniżej), nie pojawia się jako rezultat interpretacji zaobserwowanych efektów działania, ale je aktywnie współorganizuje.

Intencja w działaniu jest stanem umysłowym o nietrywialnej strukturze i trudnej do uchwycenia funkcji. Z jednej strony, filozoficzna analiza pojęciowa Searle'a wyraźnie wskazuje na przyczynową sprawczość tego typu stanu, z drugiej strony, dane eksperymentalne dla prostych intencjonalnych działań spontanicznych wskazują na jej wyłącznie korelacyjny charakter. Natomiast w odniesieniu do treści intencji obydwa typy analiz prowadzą do podobnych wniosków. Zdaniem Searle'a, intencja w działaniu funkcjonuje jako stan pośredniczący pomiędzy prior intencją (planem) a zachowaniem. Do podobnych wniosków doszli psychologowie intencji (Haggard, 2005; Pockett i in., 2006; Sirigu i in., 2003), którzy wyróżniają w tego typu intencji (w działaniu) komponent motoryczny oraz sensorycznie nacechowane odniesienie do celu działania. Największa zatem różnica między ujęciem intencji wywodzącym się z psychologii a analizą filozoficzną Searle'a dotyczy struktury i funkcji intencji. Gdy uwzględni się fakt, że nad celowościowym aspektem intencji (drugą składową treściową) nadbudowany jest jeszcze jeden fenomen, tzw. poczucie sprawstwa, to opisany problem jeszcze bardziej się komplikuje. Poniżej omawiam dokładniej poczucie sprawstwa, które jest umysłową „projekcją" sprawstwa faktycznego.

4.2 Poczucie sprawstwa

Daniel Wegner w swojej szeroko dyskutowanej książce *The illusion of conscious will* przedstawił następujący eksperyment myślowy:

> *Wyobraź sobie, że dysponujesz magiczną zdolnością przewidywania kierunku, w jakim wiatr poruszy gałęzią obserwowanego przez ciebie drzewa. Tuż przed jej poruszeniem wiedziałbyś, że za chwilę gałąź się poruszy, znałbyś kierunek ruchu – po prostu wiedziałbyś, co się stanie. Ta sama magiczna zdolność nie tylko udostępniałaby ci wiedzę o przewidywanym ruchu gałęzi, ale również gwarantowałaby, że tuż przed tego typu zdarzeniem miałbyś odpowiadającą mu myśl.*

W takim układzie patrzyłbyś na gałąź, wiedziałbyś, że za chwilę się poruszy i po chwili będziesz obserwował jej ruch.[55]

Tego typu zdolność, zdaniem Wegnera, po pewnym czasie doprowadziłaby do przekonania, że nasze myśli przyczynowo oddziałują na obserwowane drzewo, gdyż to my jesteśmy tak naprawdę sprawcami ruchów jego gałęzi, a nasza wola steruje tym, co się dzieje. Powyższy przykład jest całkowicie wyimaginowany, zawiera jednak w sobie element, który brzmi wiarygodnie, mianowicie – bez względu na to, czy jesteśmy źródłem danej zmiany w świecie, czy nie – nasz umysł w określonych warunkach zacznie postrzegać tę zmianę tak, jakbyśmy faktycznie byli jej sprawcami. W psychologii intencji tego typu fenomen zwykło się nazywać poczuciem sprawstwa (*the sense of agency*). Na poziomie treści odpowiada mu przeświadczenie, że „ja" kontroluje określone zdarzenia w świecie (Haggard, 2005, s. 293). Z kolei na poziomie fenomenalnym, wskazany stan to tzw. emocja tła, która wyraźnie ujawnia się tylko w przypadku dysocjacji.

Wystąpienie dysocjacji, czyli zaistnienie działania bez towarzyszącego mu poczucia sprawstwa prowadzi – z perspektywy agenta – do powstania przekonania, że tego typu zachowanie wywołane zostało przez czynniki zewnętrzne, na które agent nie miał wpływu. Badacze od wielu lat próbują wyjaśnić tworzenie się i funkcjonowanie mechanizmu odpowiedzialnego za pojawianie się lub nie poczucia sprawstwa. Zgromadzono szereg danych empirycznych, które wstępnie pozwalają określić najważniejsze czynniki wpływające na powstanie tego fenomenu. Zanim jednak zostaną one przedstawione, dokonany zostanie krótki przegląd najciekawszych przypadków, w których nie pojawia się poczucie sprawstwa z przyczyn neurologicznych (Bayne, 2006; Frith, 2012; Haggard, 2005).

Neurologiczne uszkodzenia mózgu – brak poczucia sprawstwa

Zdaniem Wegnera, „wyodrębnienie" poczucia sprawstwa z procesu konstytuującego przebieg działania, pozwala nieco inaczej spojrzeć na następujące schorzenia: parkinsonizm, pląsawicę czy zespół Turrette'a. W każdym z tych przypadków mamy do czynienia z działaniami, które traktowane są przez ich sprawców jako niechciane

[55] "Imagine for a moment that by some magical process you could always know when a particular tree branch would move in the wind. Just before it moved, you would know it was going to move, in which direction, and just how it would do it. Not only would you know this, but let's assume that the same magic would guarantee that you would happen to be thinking about the branch just before each move." (Wegner, 2002, s. 63).

(szczególnie widoczne jest to w przypadku zespołu Turrette'a). Osoby cierpiące na to schorzenie mają wiele nieskoordynowanych tików, wbrew własnej woli wykrzykują obraźliwe treści pod adresem towarzyszących im osób, itp. Jest to przykład pokazujący, że poczucie sprawstwa nie jest obecne w tego typu działaniach.

Innym przykładem braku poczucia dobrowolności, przywołanym przez Wegnera, jest tzw. zespół obcej ręki. Osoby dotknięte tym neuropsychologicznym zaburzeniem (współcześnie łączonym z uszkodzeniem środkowej części płata czołowego) postrzegają jedną ze swoich rąk jako całkowicie niezależną, działającą wbrew ich świadomym intencjom i celom. Wegner opisuje przypadek mężczyzny, któremu „obca ręka" ciągle przeszkadzała podczas gry w warcaby, wykonując błędne i niechciane przez niego ruchy. Potrafiła „złośliwie" zamykać właśnie otwartą książkę czy ogólnie wykonywać działania przeciwne do tych, które wykonywała ręka podlegająca świadomej kontroli. Podobne problemy miała kobieta, której „obca ręka" zabierała i chowała wszystkie blisko położone przedmioty, a podczas snu chwytała ją za gardło. Działania „obcej ręki" były traktowane przez każdą z tych osób jako niezależne od ich woli. Wymienione przypadki prowadzą do następującego wniosku: poczucie sprawstwa to niezwykle istotny element procesu realizacji działań. Jego brak prowadzi do poważnych zaburzeń i w istotny sposób wpływa na poczucie kontroli działań.

Poczucie sprawstwa a typy działań

Daniel Wegner w *The illusion of conscious will* zidentyfikował i skategoryzował relacje, jakie mogą zachodzić między różnymi typami działań a tzw. poczuciem sprawstwa. Poniższa tabela przedstawia przypadki zidentyfikowane przez Wegnera[56]:

	Poczucie sprawstwa	**Brak poczucia sprawstwa**
Działanie	Działania dowolne (intencjonalne)	Automatyzmy
Brak działania	Iluzja kontroli	Stan braku działania

Rys. 4. Relacja między poczuciem sprawstwa a typami działań (Wegner, 2002, s. 8).

[56] Należy zauważyć, że Daniel Wegner jest niekonsekwentny w używaniu terminologii. Fenomen sprawstwa często zastępuje innymi pojęciami: „poczuciem działania" (*the experience of doing*), „poczuciem wysiłku" (*the experience of effort*), „przeżyciem świadomej woli" (*experience of conscious will*), „doświadczeniem poczucia wolnej woli" (*experience of experience of free will*). Tę niekonsekwencję zarzucił mu m.in. Tim Bayne (Bayne, 2006).

Działania intencjonalne, które – wg schematu Searle'a (patrz Diagram 3.) – składają się z intencji w działaniu (czasami poprzedzonej prior intencją) oraz z jednego lub więcej zachowań (np. ruchów ciała lub wypowiedzi) poczucie sprawstwa. Poczucie to wydaje się być nadbudowane nad związkiem między intencją a zachowaniem. Introspekcja upewnia nas, że związek ten ma charakter przyczynowy, tymczasem, jak twierdzi Wegner, nie mamy podstaw, by sądzić, że związek istniejący między intencją a zachowaniem ma charakter kauzalny. Zdaniem amerykańskiego psychologa, przeświadczenie, że nasze skuteczne funkcjonowanie w świecie polega na umiejętnym wyborze zamierzeń, które z kolei inicjują działania prowadzące do pożądanych zmian, należy do twierdzeń z obszaru tzw. „psychologii ludowej" (*folk psychology*) (Hirschfeld, 2007), opierającej się na potocznych skojarzeniach i wyobrażeniach, które w konfrontacji z wiedzą naukową okazują się daleko posuniętymi uproszczeniami, a czasami nawet przekonaniami całkowicie fałszywymi. W tym kontekście Wegner proponuje, żeby najpierw przyjrzeć się przypadkom brzegowym, tj. automatyzmom, iluzji kontroli oraz wybranym przypadkom różnego rodzaju lezji (np. zachowanie używające, *utilization behavior*, które jest skutkiem uszkodzenia w płacie czołowym) a następnie, na ich podstawie, skonstruować model wyjaśniający związek między zachowaniem, intencją oraz sprawstwem. W kolejnych sekcjach tego rozdziału przedstawione zostaną najważniejsze wnioski płynące z przeprowadzonej przez Wegnera analizy, która dotyczy wskazanych przypadków. Opracowany na ich podstawie model w istotny sposób zakwestionuje oparte na introspekcji wyobrażenia odnoszące się do przyczynowej roli intencji w działaniu.

Automatyzmy – zanik poczucia sprawstwa

Należy na wstępie zaznaczyć, by uniknąć nieporozumień, że Wegner, mówiąc o automatyzmach, nie ma na myśli tzw. „zachowań automatycznych", tj. zachowań, które cechuje niekontrolowalność, brak umyślności, wysoka efektywność oraz realizacja bez aktywnego udziału świadomości (Wegner, 2002, s. 9). Autor *The Illusion of Conscious Will*, nawiązując do pracy Hartley'a z 1749 roku *Observations on man, his frame, his duty, and his expectations*, automatyzmami nazywa następujące typy zachowań: mimowolne ruchy przy stole spirytystycznym, „automatyczne pisanie" (efekt częściowo zbliżony do syndromu obcej ręki, pojawiający się nawet przy częściowym znieczuleniu[57]),

[57] Interesujący efekt pojawia się w sytuacji związanej z automatycznym pisaniem. W przeprowadzonych przez Wegnera testach, 1/3 osób skłonna była napisać na kartce pomyślane przez siebie imię, raportując równocześnie, że autorem napisu był eksperymentator. Trik, zdaniem amerykańskiego psychologa, polega

posługiwanie się planszą OUIJA[58], wahadełkiem oraz różdżką. Wymienione przypadki to przykłady zachowań, w których podmiot realizuje pewne działanie, pojmowane i postrzegane przez niego zazwyczaj jako dowolne, ale nie postrzega siebie jako jego sprawcy. Efekt ten, jak twierdzi Wegner, można wyjaśnić za pomocą teorii czynności ideomotorycznych zaproponowanej przez Williama Carpentera (Wegner, 2002, s. 99). W szczególnych warunkach, zgodnie z tą teorią, bezpośrednią przyczyną działania może być treść pewnego wyobrażenia, a nie intencja, jak to zazwyczaj bywa w takich sytuacjach. Tego typu wyobrażenie wywołuje pobudzenie określonych motoneuronów, które prowadzą do niewielkich ruchów mięśni. Można, wykorzystując specjalne urządzenie (tzw. automatograf) w połączeniu z określonymi wyobrażeniami, uchwycić efekt „sterowania" mikroskurczami. W ten sposób udaje się na przykład odczytać położenie obiektu w pomieszczeniu – wystarczy tylko poprosić uczestnika eksperymentu, aby, trzymając rękę na automatografie, myślał o miejscu, w którym obiekt został ukryty. Czułe na najmniejszy ruch urządzenie stopniowo wyznaczy ślad, z którego bez trudu będzie można określić przybliżone położenie wyobrażonego obiektu. Interesujące w tej teorii jest założenie, że myśl sterująca skurczami mięśni nie jest traktowana przez sprawcę zachowania jako ich przyczyna, że nie jest intencją.

W związku z tym powstaje pytanie: dlaczego tylko w tak szczególnych warunkach nasze zachowania są bezpośrednio wywoływane przez nasze myśli-wyobrażenia? Odpowiedź Wegnera jest następująca: gdybyśmy natychmiast realizowali zachowania pod wpływem zjawiających się myśli-wyobrażeń, bylibyśmy bardzo nieefektywni (np. wyobrażenie o maszerującej kaczce mogłoby prowadzić do zachowań naśladowczych, dysfunkcjonalnych z punktu widzenia społeczności ludzkich). Istnieje jednak takie uszkodzenie płatów czołowych, które powoduje, że mechanizm hamujący natychmiastowe reakcje na pojawiające się reprezentacje obiektów dostrzeżonych w środowisku przestaje działać. Osoba dotknięta tego typu schorzeniem podlega tzw. *utilization behavior*, czyli bezzwłocznej reakcji czynnościowej na zjawiający się w jej otoczeniu obiekt lub jego cechę. Przykładowo, jeśli osobie dotkniętej tego typu dysfunkcją podana zostanie karafka oraz szklanka bez jakiegokolwiek komentarza, to osoba dotknięta *utilization behavior*

na odpowiednim skoordynowaniu następujących czynności: najpierw osoba poddana testowi powinna odwrócić wzrok od kartki, następnie testujący powinien delikatnie położyć własną dłoń na dłoni osoby testowanej, wreszcie osoba manipulująca powinna zacząć delikatnie poruszać ręką osoby testowanej (Wegner, 2002, s. 107).

[58] Plansza OUIJA zawiera nadrukowane litery alfabetu oraz cyfry. Przy jej pomocy odbywała się „komunikacja" z duchami osób, które zostały przywołane podczas seansu spirytystycznego.

natychmiast naleje wody z karafki do szklanki, nie pytając nawet, dlaczego te przedmioty zostały jej podane. Podobny efekt pojawi się, gdy poda się choremu okulary. Nawet, jeśli badany ma już okulary na głowie, to po podaniu kolejnej pary podjęta zostanie próba założenia drugiej pary. Oznacza to, że osoby dotknięte *utilization behavior* – pod wpływem bodźców – działają w pełni mechanicznie, bez odwoływania się do jakichkolwiek myśli-intencji.

Automatyzmy, zdaniem Wegnera, odsłaniają sposoby sterowania naszymi działaniami za pomocą myśli. Wyróżnia on dwie grupy przypadków: (1) takie, w których mamy do czynienia z pewnym wyraźnym nastawieniem („pełną wyczekiwania uwagą" – seans spirytystyczny) oraz (2) przypadki, w których myśli sterujące działaniem są nieświadome (efekty wywoływane prymowaniem). Myśli w tych stanach umysłu krążą na tyle intensywnie wokół działania, że w końcu je wywołują (wprawiają w ruch stół spirytystyczny lub wykreślają trajektorię na automatografie), same jednak nie uzyskują statusu intencji, a co za tym idzie – nie są traktowane jako przyczyny działania.

Z przedstawionych analiz wynika, że działania ideomotoryczne wyróżnia brak jawnej intencji w działaniu oraz brak poczucia sprawstwa. Wykonywane mikroskurcze odpowiadające za ruch automatografu, choć sterowane myślą, nie są odbierane jako dowolne, a dodatkowo podmiot działający ma trudność z określeniem sprawcy odpowiedzialnego za ich pojawienie się. Prinz twierdzi, że czynności te są realizowane przez wyspecjalizowany system kontroli zachowań, który pomija intencje i powoduje, że działanie jest realizowane bezpośrednio na podstawie treści myśli (Prinz, 1987, s. 47–76).

Wegner zaproponował jednak inne spojrzenie na ten problem. Zasugerował, aby potraktować automatyzmy jako działania, którym brakuje fazy interpretacji, w wyniku czego nie powstaje poczucie sprawstwa. W jego opinii, takie ujęcie daje lepszy dostęp do związku między myślą a działaniem w akcie wolicjonalnym. W tym kontekście pojawia się również szczególnego rodzaju niezależność procesu interpretacyjnego od faktycznej relacji, która w danym przypadku zachodzi między myślą a zachowaniem. Niezależność tę wyraźnie widać w przypadku praktyk różdżkarskich stosowanych w związku z poszukiwaniem tzw. żył wodnych. Z systematycznych badań (Vogt, 2000) tego zjawiska wynika, że mamy tu do czynienia z typowymi automatyzmami realizowanymi za pomocą mikroskurczy. Widoczne w ich wyniku ruchy wahadełka czy różdżki są na tyle nieprzewidywalne, że poszukiwacz wody odnosi wrażenie, jakoby działały tu niezależne

siły. Innymi słowy, intencja – by znaleźć wodę – nie jest postrzegana ani jako źródło zachowania, ani jako element odpowiedzialny za poczucie sprawstwa.

Wegner zwraca również uwagę na to, że osoby doświadczające automatyzmów interpretują je jako wpływ czegoś zewnętrznego – ducha lub oddziałującej na nie bioenergii. W tej interpretacji, zdaniem amerykańskiego psychologa, artykułowana jest pewna bardziej podstawowa potrzeba, a mianowicie: chęć wyjaśnienia przyczyn danego zjawiska lub zachowania. Jeśli jakiemuś działaniu nie towarzyszy poczucie sprawstwa (przeżycie świadomej woli), nie został zidentyfikowany związek przyczynowy między myślą towarzyszącą działaniu a działaniem, to podjęta zostaje próba zidentyfikowania innego agenta, któremu można by przypisać jego autorstwo.

> *Niebywale silna korelacja między automatyzmem a przypisywaniem tego typu działań zewnętrznemu agentowi [np. duchowi] sugeruje, że w momencie, gdy widzimy zachowanie, natychmiast zakładamy, że ktoś je wywołał. […] W tym kontekście ważne jest, aby uznać prosty fakt, że działania stanowczo „domagają" się wyjaśnienia ze względu na ich sprawcę. Agenta możemy odnaleźć w sobie wtedy, gdy pojawia się iluzja świadomej woli, gdy ta iluzja się rozpada, wówczas znajdujemy go gdzieś indziej. A obecność innego potencjalnego agenta, innego niż my sami, może uwolnić nas od iluzji, że świadomie chcieliśmy zrealizować dane działanie.*[59]

Przedstawiona przez Wegnera interpretacja złożonych zachowań kulturowych, choć interesująca, jest zarazem kontrowersyjna. Tłumaczy ona – poprzez niskopoziomowy mechanizm konstrukcji poczucia sprawstwa – skomplikowane, zakorzenione w rozbudowanych systemach przekonań religijnych zjawiska/doświadczenia (Otto, 2000; Węcławski, 1995). O ile można zgodzić się z Wegnerem, że rozumienie i znajdowanie racji dla działań własnych oraz innych agentów stanowi jedną z podstawowych potrzeb i funkcji umysłu ludzkiego, o tyle problematyczne wydaje się redukowanie wpływu rozbudowanej sieci przekonań religijnych do swoistej „nadbudowy" dla błędnych atrybucji dotyczących sprawcy działania. Nietrudno sobie wyobrazić alternatywne wyjaśnienie, mianowicie, że w tego typu przypadkach dochodzi do swoistej reinterpretacji własnych zachowań, innymi

[59] "The remarkably strong link between automatism and the attribution of outside agency suggests that when we see an action, we immediately require that someone did it. […] For now, it is important to recognize just the basic fact that actions cry out for explanation in terms of an agent. That agent can be found in the self when there is an illusion of conscious will, and elsewhere when the illusion breaks down. And the presence of any potential agent other than self can relieve us of the illusion that we consciously willed our action." (Wegner, 2002, s. 143).

słowy: sprawca doskonale wie, co robi, jednak – ze względu na kontekst kulturowy i silną presję społeczną – zawiesza „standardową" interpretację i „tworzy" wyjaśnienie oparte na idei „nawiedzenia", „jedności z bóstwem", „opętania" albo innych formach religijnego doświadczenia. Przedstawiona wątpliwość, istotna z perspektywy badań religioznawczych i antropologicznych, nie kwestionuje równocześnie zasadniczej obserwacji Wegnera, mianowicie, że interpretacja własnego oraz cudzego sprawstwa jest procesem złożonym, zależnym od kontekstu i polegającym niekiedy na projekcji działania na innego agenta.

Iluzja kontroli – autonomia sprawstwa

Proces interpretacji zachowań odpowiedzialny za powstanie poczucia sprawstwa charakteryzuje się, w opinii Wegnera, daleko posuniętą autonomią. Dobrym tego przykładem są działania klasyfikowane przez amerykańskiego psychologa jako iluzja kontroli. Pojawiają się one w szczególnych okolicznościach i w pewnym sensie można je traktować jako błąd. Egzemplifikacją tego typu działania może być sytuacja, gdy – stojąc przed konsolą gier wideo i manipulując joystickiem – odnosimy wrażenie, że kontrolujemy przebieg gry; tymczasem, kiedy wzrasta nasze zaangażowanie, rozgrywka niespodziewanie się kończy. Nagle dociera do nas, że byliśmy jedynie świadkami symulacji (tzw. demo), a nasze ruchy drążkiem nie sterowały tak naprawdę niczym. Przedstawiona sytuacja wyraźnie wskazuje, że po spełnieniu pewnych warunków (patrz poniżej: zasada priorytetu, spójności i wyłączności) nawet zachowanie niezwiązane z obserwowanymi efektami może zostać uznane za kontrolowane w pełni.

Post-rekonstruktywistyczny model sprawstwa Daniela Wegnera

Daniel Wegner na podstawie m.in. powyższych danych zaproponował model działania dowolnego (intencjonalnego), w kontekście którego umieścił proces odpowiedzialny za utworzenie poczucia sprawstwa. Najważniejsze elementy modelu prezentuje poniższy diagram:

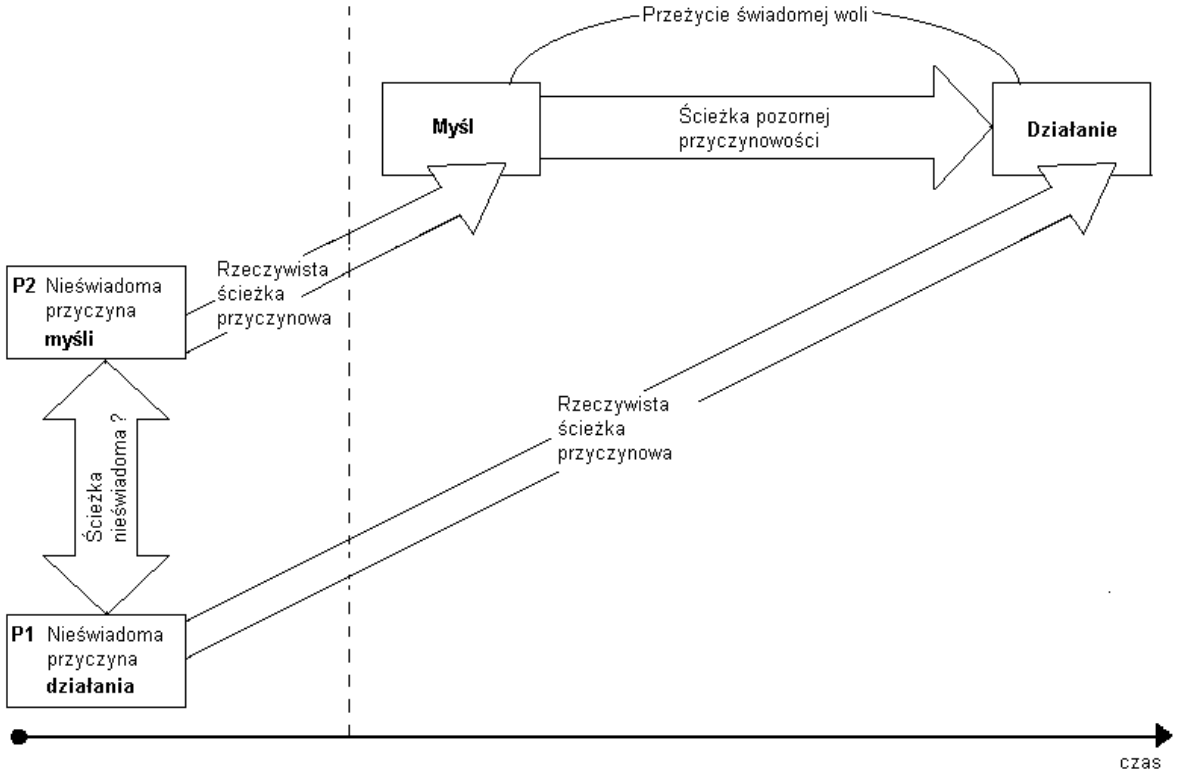


Diagram 7 Model działania dowolnego wg Daniela Wegnera.

Warto od razu zauważyć, że zaprezentowane na powyższym diagramie składowe działania dowolnego układają się na osi czasu zgodnie z przebiegiem zdarzeń neuronalno-mentalnych zarejestrowanych przez Libeta w jego, omówionym powyżej, eksperymencie. Znaczy to, że reprezentowany na diagramie akt wolicjonalny rozpoczyna się od fazy nieświadomej – dowodem jej wystąpienia jest potencjał gotowości (RP) rejestrowany około 300 ms przed pojawieniem się świadomej chęci wykonania ruchu. W fazie tej Wegner wyróżnia dwa procesy. Pierwszy z nich (P1) odpowiedzialny jest za wyznaczenie **zachowania**, drugi (P2) za „przygotowanie" **myśli** w jakiś sposób powiązanej z zachowaniem. Obydwa procesy są nieświadome, dlatego trudne jest określenie związku, który je łączy. Reprezentacje myśli i zachowania podlegają dalszemu „przetworzeniu" w procesie interpretacji, którego ostatecznym celem jest wyjaśnienie realizowanego działania w kategoriach mentalnej przyczynowości. Jeśli proces przebiegnie zgodnie z określonymi regułami, to w działanie włączone zostanie poczucie sprawstwa, swoisty „marker" wskazujący podmiotowi, że to on jest autorem zrealizowanego działania. Wegner klasyfikuje to poczucie jako tzw. uczucie kognitywne, czyli stan złożony z dwóch

elementów: (1) emocji wyrażającej się w charakterystycznym pobudzeniu somatycznym[60] oraz (2) specyficznej treści dostępnej w introspekcji.

W przypadku działań intencjonalnych treść tę można sformułować następująco: podjęte działanie realizowane jest w związku z wyznaczonym przeze mnie zamiarem jego wykonania (myślą o tym działaniu) – jest to zatem szczególnego rodzaju stan intencjonalny S(p), który posiada swoje modi psychologiczne oraz warunki spełniania. Poczucie sprawstwa jest przeżyciem stopniowalnym. W zależności od zaistniałych warunków bywa mniej lub bardziej wyraźne. Decyduje o tym treść myśli, odstęp czasowy między myślą a występującym po niej zachowaniem oraz okoliczności zewnętrzne towarzyszące działaniu. Wegner przedstawił wskazane determinanty w formie trzech zasad: zasady priorytetu, zasady spójności oraz zasady wyłączności.

1. **Zasada priorytetu** określa wielkość okna czasowego, w którym muszą zmieścić się myśl oraz działanie, by zostały uznane za powiązane relacją przyczynową. Jeśli myśl znacznie poprzedzi działanie lub przeciwnie – nastąpi zbyt późno względem działania, wówczas nie zostanie ona zinterpretowana jako jego przyczyna.

   Potwierdzają to, zdaniem Wegnera, choćby eksperymenty Michotte'a (1954) z układami poruszających się obiektów. Belgijskiemu badaczowi udało się wykazać, że obserwator klasyfikuje przemieszczające się dwie figury jako niezależne lub oddziałujące na siebie w trybie przyczynowo-skutkowym w zależności od istniejących między nimi relacji czasowo-przestrzennych. Okazało się, że krytyczny dla interpretacji przyczynowo-skutkowej jest moment, w którym rozpoczyna się ruch obiektu reprezentującego skutek – nie może on być inicjowany zbyt wcześnie ani zbyt późno. W przypadku aktów wolicjonalnych ocenia się, że myśl związana z działaniem powinna wyprzedzać je o kilka sekund, by mogło ono zostać uznane za intencjonalne. Oszacowania dopuszczają przedział od 3 sekund (efekt przełączania w świadomości kostki Neckera) do 30 sekund (górna wartość ustalona jest na podstawie zasad funkcjonowania pamięci krótkotrwałej). W wyjątkowych sytuacjach zasada priorytetu może być naruszona, tzn. myśl o działaniu może pojawić się po działaniu, a mimo wszystko zostanie ono uznane za intencjonalne. W takim przypadku źródłem poczucia świadomej woli są dwie pozostałe zasady: spójności i wyłączności.

[60] Por. Koncepcja somatycznego markera Antanio Damasio(por. Wegner, 2002, s. 326).

2. **Zasada spójności** wymaga, aby przyczynie odpowiadał adekwatny do niej skutek. W kontekście pozornego mentalnego prawa przyczynowości, konstruującego poczucie woli, zasada spójności ujawnia się poprzez ścisły związek treści myśli z podejmowanym działaniem. W praktyce, zasada ta jest spełniona dzięki temu, że w myśli towarzyszącej działaniu występuje jego nazwa lub obraz, albo określenie odnoszące się do sposobu wykonania danej czynności lub jej skutku. Znaczy to, że między daną myślą (dostępną nam w formie intencji, przekonania czy pragnienia) a działaniem pojawia się relacja o charakterze semantycznym. Spójność myśli i działań zależy od procesu poznawczego, w trakcie którego następuje porównanie wcześniejszej myśli dotyczącej działania z jego efektami. Gdy rezultat zachowania jest zgodny z utworzoną wcześniej intencją, wówczas przeżycie świadomej woli zyskuje na wyrazistości i jednoznaczności. W tym kontekście Wegner przywołuje szereg badań wskazujących, że osoby, które np. zakładają, że odniosą sukces (zamiar zostanie zrealizowany), po jego osiągnięciu mają większe poczucie kontroli nad własnym działaniem, niż te, którym również przytrafił się sukces, ale w niego nie wierzyły (np. osoby cierpiące na depresję (Wegner, 2002, s. 80)).

   Zasada spójności (w pewnym zakresie) wyjaśnia również „efekt eureki". Kiedy zjawia się nowatorska idea, to nie traktujemy jej na ogół jako przeżycia intencjonalnego – w tym przypadku poczucie sprawstwa jest niezwykle słabe. Na ogół tego typu „zaskakujące" pomysły przypisujemy nieświadomości czy wyższej istocie. Tak Jules-Henri Poincaré zinterpretował własne odkrycia dotyczące tzw. funkcji Fuchsa (Poincaré, 1914, s. 52). Francuski matematyk po wielodniowym, acz tylko częściowo skutecznym namyśle, postanowił przerwać pracę nad badanym problemem. By oderwać się od badań, zdecydował się dołączyć do grupy znajomych przebywających w Coutances. By umilić sobie czas ze znajomymi, matematyk postanowił wybrać się na przejażdżkę. W chwili, gdy wsiadał do autobusu, nagle, bez związku z bieżącymi myślami, pojawiło się rozwiązanie, które w istotny sposób posunęło do przodu prowadzone badania nad funkcją Fuchsa. Poincaré, zastanawiając się nad przebiegiem procesu twórczego, którego był zarazem uczestnikiem, doszedł do wniosku, że uzyskanie wyniku możliwe było wyłącznie na skutek nieświadomych procesów umysłowych (Falk, 2005). Podobny, do pewnego stopnia, efekt pojawia się w kontekście działań eksperckich, których rezultaty zaskakują samych twórców, np. improwizacje wybitnych muzyków często wprawiają ich samych w zdumienie. Można powiedzieć, odnosząc się do zasady

spójności, że tego typu wyjątkowe działania paradoksalnie nie są traktowane przez ich sprawców jako zamierzone, trudno jest bowiem ich autorom uspójnić uzyskane rezultaty zachowań z towarzyszącymi im myślami.

Innym przykładem niespójności myśli z działaniem jest zjawisko „słyszenia głosów" występujące m.in. u osób chorych na schizofrenię. Wegner, powołując się na pracę Hoffmana *Verbal hallucinations and language production processes in schizophrenia* (Hoffman, 1986), przywołuje następującą hipotezę: zwykle, gdy zaczynamy mówić, generujemy kognitywny, ale zarazem abstrakcyjny plan wypowiedzi. W planie tym określa się jej intencję, istotę lub cel. Nie jest to oczywiście zamknięta całość, gdyż plan jest wrażliwy na kontekst i przekonania mówcy. W trakcie wypowiedzi przekłada się go na określone słowa oraz składnię. W przypadku chorych na schizofrenię, zdaniem Hoffmana, ten plan rozsypuje się, co może prowadzić do niespodziewanych wypowiedzi i myśli. Często takie osoby twierdzą, że to, co wypowiedziały – było niezgodne z tym, co powiedzieć chciały i stąd naturalna tendencja, by wygłoszoną wypowiedź przypisać jakiemuś obcemu głosowi. Jeszcze bardziej dotkliwe są dla osoby chorej sytuacje, gdy czuje się ona zmuszona do wygłaszania sądów, które nie dość, że są niezgodne z jej własnymi intencjami, to są jej narzucone przez słyszany przez nią głos lub głosy (Hoffman, 1986).

Przytoczone przykłady wskazują na związki między sprawstwem a działaniem oraz myślami towarzyszącymi działaniu. Z zasady spójności wynika zatem, że proces wyznaczenia poczucia sprawstwa dla danego działania wymaga od agenta przeprowadzenia niejawnego wnioskowania, przy pomocy którego agent „oszacuje", na ile określone stany intencjonalne – m.in. zamiary, przekonania i pragnienia – doprowadziły do powstania obserwowanych rezultatów działania. Jeśli okażą się one niespójne, wówczas poczucie sprawstwa nie ujawni się, jak w przypadku osób cierpiących na schizofrenię i doświadczających werbalnych halucynacji, lub ujawni się częściowo, jak to się dzieje w sytuacji, gdy doświadczamy efektu eureka.

3. **Zasada wyłączności** wskazuje na to, że warunkiem powstania poczucia sprawstwa jest możliwość jednoznacznej identyfikacji przyczyny. Im więcej potencjalnych przyczyn działania, tym mniejsza możliwość zaistnienia poczucia kontroli. Czynniki obniżające poziom wyłączności mogą mieć charakter wewnętrzny albo zewnętrzny. Pierwszy rodzaj stanowią emocje, impulsy, przyzwyczajenia,

usposobienie oraz tiki nerwowe – te stany umysłowe są silnie powiązane ze stanami ciała i rywalizują z myślą o staniu się przyczyną zachowania. Przykładowo, kiedy działamy pod wpływem silnej emocji, wszelkie świadome myśli są na tyle wytłumione, że nie odbieramy naszych działań jako dowolnych. Myśl nie jest dla nas ani przymusowa, ani imperatywna. Jeśli się pojawi, to możemy od niej odwrócić naszą uwagę i pomyśleć o czymś innym. Jeśli myśl zawiera nakaz, to również nie musimy go spełniać. Natomiast emocja (podobnie jak pozostałe, wymienione wyżej stany) jest zarówno kompulsywna, jak i imperatywna. Jeśli już się pojawi, to nie możemy się od niej uwolnić. Jak wiadomo, z emocją wiąże się określone zachowanie (ucieczka, walka, cielesne „zastygnięcie", itp.), jednak to nie my decydujemy czy i jakie zachowanie wykonać, emocja „podejmuje tę decyzję" za nas, a my ją tylko wykonujemy.

> *Wola może zostać zniekształcona w wyniku rywalizacji o miano prawdopodobnego źródła danego zachowania pomiędzy myślą a wewnętrznymi zmiennymi psychologicznymi, takimi jak emocje, impulsy czy nawyki. Z kolei absencja takich wewnętrznych zmiennych może wzmocnić przypisanie działania myślom pojawiającym się w jego kontekście.*[61]

Do czynników zewnętrznych naruszających zasadę wyłączności zalicza się wpływ, który inne osoby lub grupy osób (inni agenci) wywierają na nas. Za każdym razem, gdy dochodzi do współdziałania kilku osób, silnie obniża się poczucie umyślności. Przykładem może być wspólny taniec, kłótnia, zapasy czy seans spirytystyczny. W każdym z tych przypadków trudno jest ich uczestnikom określić, które dokładnie działanie było przez nich zamierzone, a które nie. Jeszcze wyraźniej efekt ten widać w sytuacjach działań grupowych – w skrajnych przypadkach dochodzi do wyodrębnienia się agenta zbiorowego[62] (Campbell, 1958). Uczestnik tego typu grupy przestaje traktować podejmowane przez siebie działania jako powiązane z jego własnymi myślami. W to miejsce wkracza agent zbiorowy: to grupa myśli o

[61] "When internal psychological variables such as emotion, impulse, and habit vie with thought as plausible sources of behavior, will can suffer as a result. The absence of such internal causes, in turn, can bolster the attribution of action to the occurrence of appropriate action relevant thoughts." (Wegner, 2002, s. 93).

[62] Proces wyłaniania się podmiotu zbiorowego od wieków intrygował twórców kultury. W ekranizacji powieści Patricka Süskinda *Pachnidło: historia mordercy* z 2006 roku w scenie, w której główny bohater ma zostać poddany egzekucji, ukazuje orgię oraz charakterystyczną dla tego typu zjawiska dynamikę, czyli fazę włączania się kolejnych jednostek w działania grupy, tymczasowe funkcjonowanie w obrębie podmiotu zbiorowego oraz moment, gdy kolejne osoby otrząsają się z sytuacji i wracają do swych społecznych ról.

czymś i to grupa podejmuje określone działanie, a nie indywiduum. Z analogicznym zjawiskiem, zdaniem Wegnera, mamy do czynienia w przypadku ekstatycznych transów, rozpowszechnionych w wielu kulturach, w trakcie których ujawniają się takie zdolności, jak przemawianie nieznanymi językami, przepowiadanie przyszłości itd.

Autor *The illusion of conscious will* zwraca również uwagę na wpływ zasady wyłączności na formowanie się indywidualnej tożsamości.

> *W rozległym polu możliwych przyczyn zachowania danej osoby istnieje tylko jedno «ja», autor, posiadający myśli i realizujący działania. To «ja» rywalizuje z przyczynami wewnętrznymi i z całą gamą przyczyn zewnętrznych o status tego, czego dana osoba naprawdę chciała. Eliminowanie po kolei wszystkich innych przyczyn działania pozwala jednostce rozwinąć to «ja» i w ten sposób doświadczyć własnej tożsamości, natomiast proces znajdowania zewnętrznych przyczyn własnych działań rozwija zdolności przypisywania woli wszystkim innym aktorom z otoczenia społecznego.*[63]

Aby przetestować powyższy model oraz powiązane z nim trzy zasady, Wegner zaproponował eksperyment, za pomocą którego postanowił zweryfikować wpływ, głównie zasady priorytetu, na percepcję sprawstwa. Zadaniem uczestników w zaaranżowanej sytuacji było raportowanie, jak bardzo realizowane w danym momencie działanie było przez nich spostrzegane jako intencjonalne. Działanie składało się z dwóch faz. W pierwszej fazie, trwającej ok. 30 sekund, badany wraz z współtowarzyszem (niejawnym przedstawicielem zespołu badawczego) proszony był, by korzystając ze specjalnie zaprojektowanej planszy (wzorowanej na planszy Ouija), w ściśle określonym czasie, wykonał koliste ruchy kursorem widocznym na ekranie monitora.

[63] „In the extensive field of possible causes of a person's behavior, there exists only one self, an author that has thoughts and does actions. This self competes with internal causes and with an array of external causes of action in the individual's assessment of what he or she has willed. Whittling away all the other possible causes of actions allows the person to develop this self and so experience personal identity, and the process of finding external causes of one's own action, in turn, gives shape and attributed will to all the other actors in one's social world." (Wegner, 2002, s. 95).

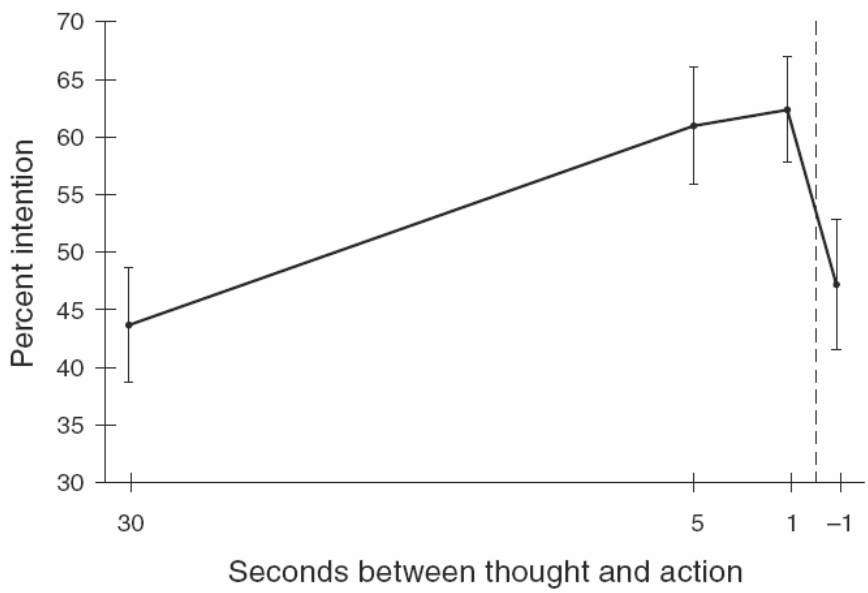
Ilustracja 1. Układ dla eksperymentu I-Spy (Wegner, 2002, s. 75).

Druga faza, ok. 10 sekundowa, polegała na przerwaniu ruchu i zatrzymaniu kursora w wybranym przez uczestników miejscu. Przejście między fazami było sygnalizowane pojawieniem się muzyki w słuchawkach. Badany był poinformowany, że w trakcie realizowania zadania usłyszy w słuchawkach pewne słowo, które będzie pełniło funkcję dystraktora. Wśród prezentowanych słów będą zarówno nazwy obrazków znajdujących się na ekranie, jak i nazwy obiektów niepowiązanych z osadzoną na pulpicie grafiką. Co szczególnie istotne, wpływ na poruszanie kursorem miał również współtowarzysz, który – w odróżnieniu od uczestnika eksperymentu – miał zawsze dążyć do postawienia kursora na ściśle określonym obrazku. Warto dodać, iż badany nie miał świadomości, że osoba siedząca po drugiej stronie planszy realizuje inne zadanie. Stworzono ciekawy układ z perspektywy badań nad działaniami dowolnymi. W punkcie wyjścia „zaburzono" bowiem zasadę wyłączności i w ten sposób utrudniono ocenę wpływu danego uczestnika na przebieg działania. Należy dodać, by zrozumieć uzyskane wyniki, że reprezentant zespołu w trakcie drugiej fazy realizacji działania otrzymywał dwojakiego rodzaju instrukcje: (1) nie działać lub (2) zatrzymać kursor na obrazku wskazanym w instrukcji.

Z perspektywy uczestnika znaczyło to, że w pewnych przypadkach kursor był pod jego całkowitą kontrolą, a w niektórych (1/4 wszystkich prób) kontrolę w dużym stopniu

sprawował reprezentant zespołu. W trakcie prób „z mocno ograniczoną kontrolą" – wbrew temu, co przekazano badanemu podczas omawiania instrukcji – pojawiały się tylko ściśle określone słowa. Były to zawsze nazwy widocznych na ekranie obrazków, na których reprezentant zespołu miał postawić kursor w drugiej fazie działania. Można powiedzieć, że w tego typu próbach uczestnik był do pewnego stopnia torowany[64] nazwą obrazka – i to właśnie wpływ torowania na percepcję sprawstwa był głównym przedmiotem eksperymentu.

Wegner zaprezentował uzyskane wyniki za pomocą następującego wykresu:



Rys. 5. Wyniki eksperymentu I-Spy (Wegner, 2002, s. 22)

Na osi rzędnych znajduje się procentowo określone poczucie umyślności (100% znaczy – „to ja zatrzymałem kursor", 0% – „to nie ja byłem sprawcą zatrzymania kursora"), natomiast na osi odciętych – czas (w sekundach), w którym prezentowane były prymy. Zaprezentowane na wykresie dane odnoszą się tylko do prób, w których działanie było „wsparte" przez manipulacje reprezentanta zespołu.

Uzyskane wyniki jednoznacznie potwierdziły wpływ zasady priorytetu na percepcję sprawstwa. Im wcześniej pryma (nazwa obrazka) słyszana była przez uczestnika w odniesieniu do docelowego położenia kursora, tym poczucie sprawstwa było mniejsze.

[64] „Torowanie, poprzedzanie (*priming*): na ogół oznacza proces prezentacji bodźca lub zdarzenia, przygotowujący system do funkcjonowania; w psychologii poznawczej: wyzwalanie określonych wspomnień za pomocą określonych wskazówek (np. „pałac" wyzwoli jedno znaczenie słowa „zamek", a „klucz" – drugie)" (Reber, 2002).

Badani mieli największe poczucie kontroli, kiedy pryma pojawiała się na 1 sekundę przed zatrzymaniem kursora. Po przekroczeniu tej granicznej wartości poczucie wpływu na zatrzymanie się kursora malało w bardzo szybkim tempie, osiągając niską wartość w chwili, kiedy nazwa pojawiała się po zatrzymaniu kursora (w chwili: 1 sekunda). Dla pełnego obrazu należy dodać, że wartość poczucia sprawstwa mierzona w próbach, w których badany miał pełną kontrolę nad kursorem, oscylowała wokół 56%. Ta stosunkowo niska wartość wskazuje, jak naruszenie zasady wyłączności wpływa na percepcję poczucia sprawstwa. Przyjmuje się zgodnie z tą zasadą, że im więcej potencjalnych przyczyn danego zachowania, tym trudniej nam ocenić nasz wpływ na jego zaistnienie.

Przedstawiony powyżej model Daniela Wegnera, prezentujący mechanizmy i zasady powstawania sprawstwa, wyraźnie wskazuje na jego rekonstruktywistyczny charakter. Model ten demonstruje, że poczucie kontroli pojawia się dopiero wówczas, gdy proces interpretacji uzyskał dostęp do:

1. składowych intencji w działaniu, w szczególności tzw. drugiej składowej, czyli odniesienia do docelowego obiektu lub zdarzenia,
2. przewidywanego lub zaobserwowanego zachowania oraz jego faktycznego rezultatu,
3. bieżącego kontekstu, na który składają się m.in. towarzyszący działaniu inni agenci.

Z pomocą pozyskanej informacji o intencji, zachowaniu, jego skutku oraz kontekście (por. punkty 1.-3.) podjęta zostaje próba powiązania tych zjawisk w formie związku przyczynowo-skutkowego. Wegner przywołuje w tym kontekście sformułowaną przez niego oraz Thalię Wheatley teorię pozornego związku przyczynowo-skutkowego (Wegner & Wheatley, 1999). „Ludzie doświadczają świadomej woli [poczucia sprawstwa], kiedy interpretują własną myśl jako przyczynę swojego działania.“[65] (Wegner, 2002, s. 64). „Wola jest doświadczana jako rezultat samopostrzegania pozornej mentalnej przyczynowości.”[66] (Wegner, 2002, s. 66). Zdaniem Wegnera, jeśli weźmiemy pod uwagę całość aktu wolicjonalnego, to przeżycie świadomej woli jest tylko kolejnym rezultatem nieświadomych procesów mózgowych oraz zdarzeń mentalnych. W efekcie powstaje, oprócz działania oraz intencji, poczucie sprawstwa, w którym intencja pojmowana jest jako

[65] *People experience conscious will when they interpret their own thought as the cause of their action* (Wegner, 2002, s. 64).

[66] *Will is experienced as the result of self-perceived apparent mental causation* (Wegner, 2002, s. 66).

przyczyna działania, choć, jak już wspomniano, nie jest to rzeczywisty związek przyczynowy, a jedynie pozorny. Warto w tym kontekście zauważyć, że w modelu Wegnera nie została zdefiniowana relacja między nieświadomym procesem odpowiedzialnym za przygotowanie zachowania (P1) a procesem odpowiedzialnym za opracowanie myśli (P2). Wegner twierdzi, że kształt tej relacji jest nieistotny z perspektywy sprawstwa.

Predykcyjny model poczucia sprawstwa

Przedstawiona powyżej koncepcja poczucia sprawstwa nie jest jedyną próbą wyjaśnienia tego fenomenu. Od ponad 20 lat rozwijany jest, oprócz propozycji Wegnera, tzw. model komparatora (Frith, 2012), który służy wyjaśnieniu zjawiska kontroli za pomocą operacji porównania dwóch stanów: (1) przewidywanych skutków działania obliczonych za pomocą modeli wyprzedzających (*forward model*) oraz (2) zaobserwowanych, rzeczywistych skutków zachowania. Jeśli w wyniku porównania okaże się, że zaobserwowany stan świata zgodny jest ze stanem przewidywanym, wówczas pojawi się poczucie sprawstwa. W przypadku niezgodności tych dwóch stanów podmiot działający uzna, że nie miał wpływu na działanie (por. Frith, 2012). Model komparatora – w odróżnieniu od modelu Wegnera – odnosi się głównie do niskopoziomowych mechanizmów predykcyjnych (głównie do przewidywanych sensorycznych skutków zachowania).

W dalszej części pracy wykorzystany zostanie model Wegnera, który umożliwia połączenie procesu konstrukcji poczucia sprawstwa z mechanizmami odpowiedzialnymi za zarządzanie siecią stanów intencjonalnych.

Dotychczasowe rozważania dotyczące intencji w działaniu oraz poczucia sprawstwa odnosiły się głównie do struktury oraz mechanizmów organizujących omówione fenomeny. Zgromadzona wiedza pokazuje, jak złożoną strukturę posiadają pozornie proste zjawiska i jak są zarządzane przez skomplikowane mechanizmy. Po dokładniejszej charakterystyce składowych prostych działań intencjonalnych można przejść do pytania o ich funkcję.

4.3 Funkcjonalne aspekty intencji oraz poczucia sprawstwa

Kiedy badacz analizuje określone zjawisko będące częścią obszerniejszej całości, jego strukturę oraz mechanizmy, często jego uwadze umyka prosta zdawałoby się kwestia: do czego tak naprawdę służy badane zjawisko, jaką funkcję pełni ono w danym systemie. To, mogłoby się wydawać, banalne pytanie nabiera szczególnej wagi w przypadku organizmów biologicznych, które, jak zauważył Darwin, poddane są ciągłej presji doboru naturalnego (Brandon, 1978). W takich warunkach u organizmów danego gatunku stopniowo zanikają cechy niekorzystne (tzn. te, które obniżają ich wartość przystosowawczą), a utrwalają i rozwijają cechy zwiększające szansę na przetrwanie. W tak określonym kontekście wcale nie jest oczywista odpowiedź na pytanie o funkcję intencji w działaniu oraz o poczucie sprawstwa. By sformułować odpowiednio problem, warto przypomnieć, że intencja w działaniu, zarówno w myśleniu potocznym (Wegner, 2002, s. 3), jak i w teorii intencjonalności Searle'a postrzegana jest jako przyczyna zachowania. Ten szczególny stan intencjonalny, zdaniem amerykańskiego filozofa, wyznacza określony „kształt" danego zachowania poprzez tzw. samo-odniesienie przyczynowe oraz odpowiednią treść. Ten stosunkowo prosty opis mocno się skomplikował w ostatnich latach za sprawą psychologów intencji oraz neuronaukowców. Intencja w działaniu okazała się zbyt prostym konstruktem teoretycznym. Wyniki eksperymentu Libeta[67] oraz szeregu innych badań neuropsychologicznych ujawniły złożoność procesów przygotowujących samo fizyczne działanie. W efekcie, intencja w działaniu została zdekomponowana na szereg składowych oraz straciła swój przyczynowy status. Warto w tym miejscu przypomnieć, podążając za Searlem, że z działaniami intencjonalnymi związane są dwa typy intencji: tzw. prior intencja (dotyczy zaplanowanych działań) oraz intencja w działaniu, która albo realizuje część lub całość planu określonego przez prior intencję, albo pojawia się w kontekście działań spontanicznych. Można zaproponować, uwzględniając powyższą dystynkcję oraz najważniejsze wyniki psychologii intencji (Haggard, 2005), taką oto konstrukcję modelu kontroli motorycznej (Haggard, 2005; Miall & Wolpert, 1996):

[67] W początkowych partiach niniejszego rozdziału omówiłem eksperyment Libeta, jego paradoksalny wynik i dyskusję, jaką wywołał.

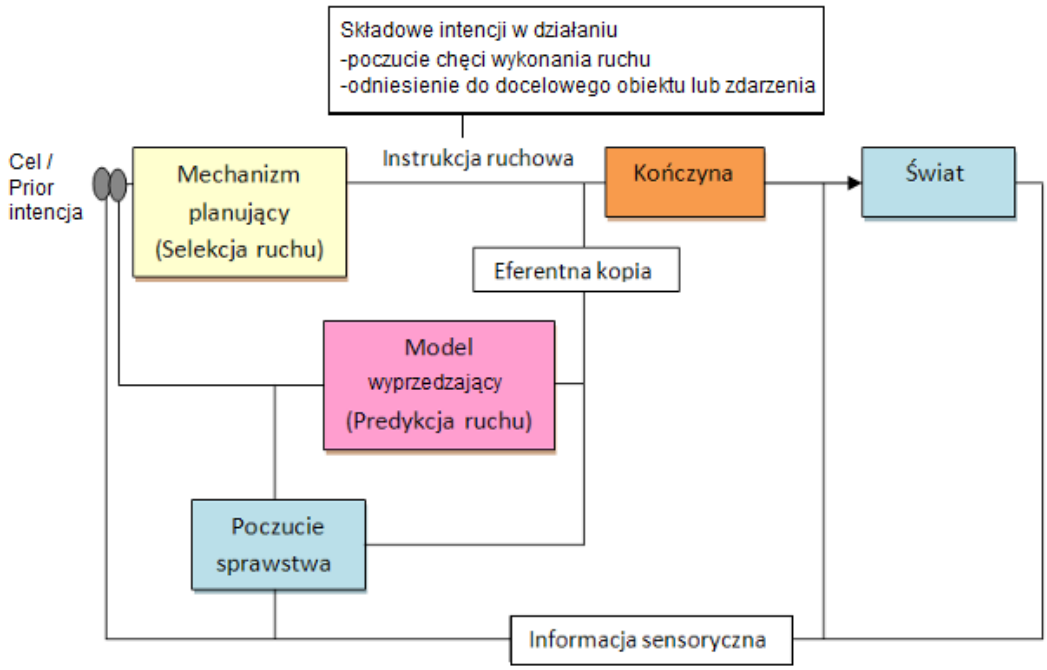


Diagram 8. Model kontroli ruchowej (por. Haggard, 2005).

Powyższy model doprecyzowuje, które miejsce w strukturze działania zajmują intencja w działaniu, jak i poczucie sprawstwa. Diagram 6. pokazuje, że intencja w działaniu korzysta głównie z informacji pochodzących ze stanu wyznaczającego cel działania (prior intencji), natomiast poczucie sprawstwa przetwarza dane pochodzące z kopii eferentnej programu motorycznego, modelu wyprzedzającego, informacji sensorycznych oraz reprezentacji celu. Widać zatem, że oba stany (intencja w działaniu oraz poczucie sprawstwa) pojawiają się na różnych etapach realizacji działania oraz korzystają z innych danych wejściowych. W związku z tym można przyjąć, że w układzie pełnią one różne funkcje, przy czym poczucie sprawstwa – zgodnie z ujęciem prospektywnym – bazuje w dużym stopniu na celowościowej składowej intencji. Wskazana zależność uzasadnia, by najpierw przeanalizować hipotezy dotyczące funkcji składowych intencji w działaniu, a następnie na tym tle zastanowić się nad rolą poczucia sprawstwa.

Funkcje intencji w działaniu

Intencja w działaniu, zgodnie z rozszerzonym przez Haggarda modelem Wolperta-Mialla dla prostych działań motorycznych, to korelat towarzyszący instrukcjom ruchowym. Korelat ten zawiera w sobie składową motoryczną (chęć wykonania ruchu) oraz składową celowościową (odniesienie do docelowego obiektu lub zdarzenia). W literaturze przedmiotu trudno jest znaleźć analizę, która w wyczerpujący sposób odnosiłaby się do

funkcji wskazanych składowych. Najbardziej wiarygodna wydaje się propozycja Patricka Haggarda, który przypisał intencji w działaniu, w szczególności składowej celowościowej, funkcję polegającą na uczeniu predykcyjnym (*predictive learning*) (Haggard, 2012a).

Haggard przytacza w tym kontekście następujący przykład: wyobraźmy sobie, że nie otrzymaliśmy od pracodawcy podwyżki z powodu niesprawiedliwej oceny okresowej wystawionej przez przełożonego. Postanawiamy, mocno rozczarowani całą sytuacją, w dosadny sposób napisać przełożonemu, co myślimy o jego postępowaniu. Bez chwili zastanowienia uruchamiamy program pocztowy i zaczynamy tworzyć „cierpki" mail. Kiedy nasz list jest już gotowy i właśnie przesuwamy myszkę, by kliknąć w przycisk „wyślij", nagle pojawia się w naszym umyśle pytanie: czy naprawdę chcesz to zrobić, czy naprawdę chcesz przekazać przełożonemu to, co znajduje się w liście? Wszystko odbywa się w ułamkach sekund i ostatecznie prowadzi do rezygnacji z wysłania listu (Haggard, 2012b). Tego typu sytuacja, z pewnością bliska wszystkim osobom zatrudnionym w korporacjach, pokazuje, zdaniem Haggarda, do czego służy intencja, w szczególności jej składowa celowościowa. Jej głównym zadaniem jest udostępnienie nam tuż przed realizacją zachowania jego przewidywanych skutków. W ten sposób podmiot działający uzyskuje możliwość stworzenia wartościowego – z perspektywy kontroli zachowań – skojarzenia: „przewidywany skutek intencji" ←→ „rzeczywisty skutek". Jeśli uzyskany rezultat okaże się dla nas niekorzystny w wyniku działania, to następnym razem intencja pozwoli nam aktywować i wyhamować tego typu działanie, a przynajmniej na chwilę powstrzymać się od jego realizacji. W artykule zatytułowanym *To Do or Not to Do: The Neural Signature of Self-Control* Haggard razem z Marcelem Brasem (2007) wykazali eksperymentalnie, że w realizację dowolnego działania zaangażowana jest grzbietowa część kory czołowo-przyśrodkowej (*dorsal fronto-median cortex*), która aktywuje się wówczas, gdy planowane działanie intencjonalne zostaje wyhamowane. Mechanizm odpowiedzialny za możliwość zawetowania aktualnie realizowanych działań dowolnych, inaczej niż twierdzi Benjamin Libet (Libet, 2004, s. 137), jest jak najbardziej uwarunkowany i nie może być argumentem potwierdzającym istnienie wolnej woli.

Warto również zauważyć, że funkcja składowej motorycznej intencji w działaniu nie została dotychczas określona. Wydaje się, że z perspektywy badań nad prostymi działaniami dowolnymi tego typu reprezentacja może stanowić co najwyżej wsparcie dla składowej celowościowej, jednak dla samej realizacji działania okazuje się ona mało istotna. Inaczej kwestia ta prezentuje się, gdy uwzględnimy kontekst poznawczy działania,

tzn. związek składowej motorycznej z mechanizmami odpowiedzialnymi za tworzenie i zarządzanie siecią stanów intencjonalnych. W takim kontekście nieodzowny jest dostęp do reprezentacji zachowań. Trudno sobie wyobrazić powstanie jakichkolwiek stanów intencjonalnych bez zdolności do reprezentowania realizowanych ruchów czy wypowiadanych słów. O reprezentacjotwórczej funkcji intencji piszę obszerniej w ostatnim rozdziale pracy.

Funkcja poczucia sprawstwa

Funkcję poczucia, że to ja wywołałem określony ruch oraz jego konsekwencje, a także kontrolowałem jego przebieg, można rozpatrywać w dwojaki sposób. W wąskim sensie – jako tzw. emocję autorstwa (*the emotion of authorship*), która zgodnie m.in. z rozszerzonym przez Haggarda modelem Wallperta-Milla opiera się w głównej mierze na przetworzeniu informacji związanych z organizacją ruchów (patrz: kopia eferentna, predykcja ruchu, informacje sensoryczne) (Braun i in., 2018; Ciechanowski, 2015; Haggard, 2005). Tak rozumiane sprawstwo pomaga nam odróżniać działania własne od działań innych agentów. Tego typu wiedza jest szczególnie przydatna w kontekstach społecznych, gdy musimy współpracować z innymi agentami. Wówczas oczekuje się od nas potwierdzenia, że realizowane przez nas działania są przez nas kontrolowane.

W szerokim sensie – poczucie sprawstwa to fenomen reprezentujący złożony proces odpowiedzialny za pogłębienie rozumienia naszych działań. Na pytanie: dlaczego świadoma wola (poczucie sprawstwa) w ogóle istnieje?, Daniel Wegner odpowiada w następujący sposób:

> *Odpowiedź na to pytanie stanie się oczywista, gdy świadomą wolę uznamy za przeżycie, które współorganizuje oraz kształtuje nasze rozumienie naszych własnych zachowań. Świadoma wola jest sygnałem przypominającym pod względem własności emocję, która przenika nasz umysł i ciało wskazując, w przypadku których działań czujemy się ich autorami. […] emocja sprawstwa realizuje kluczową funkcję w dziedzinach dotyczących naszych dokonań oraz moralności. Wydaje się, że poczucie, że coś właśnie robimy, jest bazą dla tego, co próbujemy osiągnąć oraz dla tego czy oceniamy własne postępowanie jako moralnie słuszne, czy też niesłuszne.*[68]

[68] "[…] why the conscious experience of will might exist at all. Why, if this experience of will is not the cause of action, would we even go to the trouble of having it? What good is an epiphenomenon? The answer becomes apparent when we appreciate conscious will as a feeling that organizes and informs our

Pogłębione rozumienie zachowań to proces realizowany w trybie ciągłym, obejmujący zarówno zachowania własne, jak i innych agentów. U podstaw tego procesu leży fundamentalna potrzeba, aby zaobserwowane zachowania/zjawiska wyjaśnić w kategoriach przyczynowo-skutkowych (Heider & Simmel, 1944). W związku z tym, zdaniem Wegnera, agent musi przeprowadzić dość złożone wnioskowanie (zazwyczaj nieświadome), które pozwoli orzec, jaka przyczyna (ew. przyczyny) wywołała dany rezultat. Rozważania Wegnera pozwalają przyjąć, że w tego typu wnioskowaniu uwzględniane są informacje o:

- rodzaju systemu, którego zachowanie wymaga wyjaśnienia (mentalna strategia wyjaśniania zachowań vs. mechanistyczna strategia wyjaśniania zjawisk (Baron-Cohen, 1995; Leslie, 1994)),
- relacji osoby wyjaśniającej zachowanie względem sprawcy działania (agent domowy *vs*. agent obcy *vs*. agent wirtualny)
- liczbie agentów zaangażowanych w działanie (agent indywidualny *vs*. agent zbiorowy)
- intencjach, planach i pragnieniach agenta (Ajzen, 1991; Smith, 2003),
- posiadanym przez agenta systemie przekonań (Goldman, 1976), o ile dysponuje on tego typu reprezentacjami (systemy sztuczne, np. roboty *vs*. zwierzęta *vs*. ludzie).

Wnioskowanie uwzględniające powyższe informacje wymaga od podmiotu posiadania odpowiednich kompetencji i doświadczeń. Znaczy to, że prawidłowe i skuteczne działanie mechanizmu odpowiedzialnego za przeprowadzanie tego typu rozumowań okupione jest szeregiem błędów popełnionych podczas długotrwałego procesu uczenia się. Wegner, powołując się na obszerny materiał empiryczny oraz wybrane teorie psychologiczne, rekonstruuje najważniejsze etapy procesu kształtowania się tego typu mechanizmu oraz charakterystyczne błędy popełniane wówczas, gdy agent nie dysponuje odpowiednimi kompetencjami poznawczymi lub musi się zmierzyć z wyjaśnieniem specyficznego zjawiska lub zachowania. Poniżej przedstawione zostaną wybrane elementy analizy Wegnera odnoszące się do wymienionych składowych postulowanego rozumowania. Ich prezentacja będzie miała na celu uwypuklenie dwóch kwestii: (1) stosowanych strategii reprezentowania zachowań/zjawisk oraz (2) rodzajów wiedzy wykorzystywanych do określenia sprawcy działania. Wybór odpowiedniej strategii jest jednym z pierwszych

understanding of our own agency. Conscious will is a signal with many of the qualities of an emotion, one that reverberates through the mind and body to indicate when we sense having authored an action. […] the emotion of authorship serves key functions in the domains of achievement and morality. It seems that the feeling that we are doing things serves as a basis for what we attempt to accomplish and how we judge ourselves to be morally right or wrong." *(Wegner, 2002, s. 318).*

wyzwań na drodze do prawidłowego wyjaśnienia jakiegoś zjawiska lub zachowania. Do dyspozycji, zdaniem psychologów, są dwie możliwości: wyjaśnienie mechanistyczne albo mentalistyczne. Na wysokim poziomie ogólności realizują one identyczny cel – pozwalają zinterpretować postrzegane zachowania/zjawiska w kategoriach przyczynowo-skutkowych. Każda z dwóch strategii stosuje jednak odmienny schemat poznawczy. Wskazać można dwie zasadnicze różnice: po pierwsze – gdzie indziej lokalizowane jest źródło przyczynowości, po drugie – stosowana jest inna wiedza wyjaśniająca dany związek przyczynowo-skutkowy. Przyczyna – w przypadku wyjaśniania mechanistycznego – ma charakter zewnętrzny i jest niezależna od skutku. Przykładem może być tocząca się kula, która zbija kręgle – w tym układzie kula (i jej własności) jest całkowicie niezależna od kręgli i ich własności. Sprawa przedstawia się inaczej w przypadku zjawisk wyjaśnianych przy wykorzystaniu strategii mentalnej. W tym kontekście przyczyna działania ma charakter endogenny – to stan umysłowy znajdujący się „wewnątrz" pewnego obiektu, który zdolny jest do wprawienia go w ruch. System, do opisu którego stosuje się strategię mentalną, by wyjaśnić zachowanie, w psychologii zwykło się nazywać agentem. Do jego konstytutywnych cech zalicza się zdolność do realizacji celów, posiadanie intencji i pragnień. Status agenta, oprócz zwierząt i ludzi, mogą w pewnych przypadkach uzyskać również złożone urządzenia techniczne, np. roboty realizujące odpowiednio skomplikowane zadania[69]. Zdolność do postrzegania zachowań określonych obiektów jako agentów wykształca się stopniowo – wraz z rozwojem ontogenetycznym, równocześnie jednak, jak twierdzi Baron-Cohen, niemal od urodzenia dysponujemy wyspecjalizowanym mentalnym modułem, tzw. detektorem intencjonalności, który pozwala noworodkom interpretować ruchome bodźce wzrokowe w kategoriach intencjonalnych (Baron-Cohen, 2004). Przykładem zachowania, do wyjaśnienia którego zastosowanie strategii mentalnej jest efektywniejsze, niż użycie strategii mechanistycznej, może być atak tygrysa podczas polowania. Przyczynę zaobserwowanego zachowania, stosując mentalną strategię wyjaśniającą, lokalizujemy wewnątrz drapieżnika i traktujemy ją jako najważniejszy, bezpośredni czynnik, który spowodował jego atak, natomiast to, czy w środowisku zaistniały jakieś zdarzenia (ewentualne pobudki typu: dostrzeżenie ofiary, jej zachowanie, itp.), ma charakter pośredni i drugorzędny. Rzeczywistą przyczynę, czyli decyzję

[69] Warto zauważyć, że czasami nawet eksperci (robotycy, informatycy) wykorzystują mentalną strategię wyjaśniającą do opisu zachowań konstruowanych przez siebie systemów. W wielu bowiem przypadkach tego typu wyjaśnienie, choć niekompletne i nieprecyzyjne, bywa użyteczne i efektywne.

zaatakowania, lokujemy w mózgu drapieżnika. Posłużenie się nieadekwatną dla danej sytuacji strategią może prowadzić do wyjaśnienia nieefektywnego czy wręcz błędnego.

Dzieci w początkowych fazach rozwoju, jak wykazały badania Jeana Piageta, często stosują nieadekwatną strategię – z punktu widzenia osoby dorosłej – do wyjaśniania danego zjawiska (Piaget, 1964). W efekcie prowadzi to do nieporozumień lub niesprawiedliwych ocen ze strony rodziców. Dość powszechnym zjawiskiem jest na przykład traktowanie przez dzieci przedmiotów martwych jako szczególnego rodzaju agentów posiadających intencje i zamiary (np. statek zabawka wypływa na powierzchnię po zanurzeniu, ponieważ wyimaginowany marynarz znajdujący się pod pokładem nie lubi przebywać pod wodą). Innym, często spotykanym, błędem popełnianym przez dzieci w wieku kilku lat jest brak uwzględniania intencji innych osób w wyjaśnianiu ich zachowań (przede wszystkim tych zachowań, które sprawiają dziecku przykrość albo są w inny sposób dolegliwe). W takich sytuacjach dla dziecka liczy się przede wszystkim skutek, a nie zamiar (np. nie jest ważne, dlaczego jedno dziecko wepchnęło drugie w kałużę, ale to, że zniszczone zostały buty) (Wegner, 2002, s. 22).

Badania nad rozwojem jawnej teorii umysłu (*explicit ToM*) wskazują, że od ok. 3 do 5 roku życia stopniowo nabywamy umiejętność prawidłowego przypisywania stanów mentalnych (przekonań, pragnień, intencji) innym osobom oraz sobie (Kulke i in., 2019).

> *Dzieci – bez w pełni rozwiniętej zdolności do mentalizacji – czasami nie potrafią przypisywać intencji tam, gdzie zwykle się to robi (np. osądzając ludzi tylko na podstawie skutków ich działań), jak również niekiedy przypisują je tam, gdzie się tego nie robi (np. traktując zachowania obiektów jako celowe). Dzieci mają problem ze zbudowaniem obrazu własnego umysłu, a także umysłów innych ludzi, jak również wytyczenia granicy między obiektami, którym w określonych warunkach przypisuje się posiadanie umysłu a tymi, których zachowania są zrozumiałe na mocy fizycznych (w potocznym sensie tego słowa) zależności przyczynowych. Wcześniej zakładają one, że rzeczy nieożywione mogą mieć intencje, czyli „umysłopodobne" właściwości, a niektóre rzeczy ożywione, o których później się dowiedzą, że posiadają umysły, mogą takiej intencji nie posiadać.*[70]

[70] "Without a fully developed idea of mental processes, children can fail to attribute intent when they should (in judging human beings) and attribute it too often when they shouldn't (in judging objects). Children are faced with the problem of building a picture of their own minds and the minds of others, and of achieving

Problem zmiany strategii wyjaśniającej dotyczy nie tylko dzieci, ale również osób dorosłych, które, zdobywając nowe doświadczenia i wiedzę, mogą zmienić wykorzystywany dotychczas sposób wyjaśniania na mechanistyczny, np. kiedy dowiedzą się o czysto fizycznym charakterze wyładowań atmosferycznych, przestaną traktować burzę z piorunami jako skutek intencjonalnego działania istoty nadprzyrodzonej. Charakterystyczne jest również to, że stosowanie strategii mentalnej do wyjaśniania zachowań przedmiotów fizycznych powoduje ich antropomorfizację (rodzaj animizmu (Mead, 1932)), natomiast stosowanie czysto mechanistycznego podejścia do ludzi lub zwierząt powoduje ich reifikację, a w konsekwencji może prowadzić do pojawienia się zaburzeń w relacjach międzyludzkich (np. autyzm (Sacks, 1999)). Dobór strategii poznawczej wpływa w znaczący sposób na rozumienie danego zjawiska.

Widać istotną różnicę, zdaniem Wegnera, gdy porównuje się mechanistyczną i mentalną strategię wyjaśniania zjawisk. W przypadku tej ostatniej nie mamy do czynienia z ustalaniem rzeczywistego związku przyczynowo-skutkowego, lecz z pewnym skrótem, który zastępuje ów związek (Wegner, 2002, s. 27–28). Powodem takiego stanu rzeczy jest złożoność „maszynerii" odpowiedzialnej za organizację naszych zachowań oraz brak dostępu do kluczowych danych (m.in. dotyczących wewnętrznego stanu agenta, wpływu wcześniejszych doświadczeń i przekonań na jego wybory, itp.), które okazują się niezbędne do jego wyjaśnienia. W konsekwencji, odmienny jest również rodzaj stosowanej wiedzy. W przypadku systemu mechanistycznego wykorzystuje się wiedzę z zakresu intuicyjnej wersji fizyki (*intuitive versions of physics*), często swobodnie odnoszącej się do fizyki naukowej, natomiast „rozumowania" odwołujące się do systemu mentalnego opierają się w głównej mierze na niejawnych teoriach psychologicznych (*implicit psychological theories*)[71]. Znaczy to, że w przypadku ludzi oraz naszych własnych zachowań uwzględniamy znacznie więcej informacji, niż w przypadku zwierząt lub robotów. Szczególnie istotne są w tym kontekście intencje, plany oraz pragnienia agenta. W najbardziej skomplikowanych sytuacjach, aby zrozumieć czyjeś zachowania, możemy dodać wiedzę na temat przekonań sprawcy oraz jego wcześniejszych doświadczeń, np.

an understanding of what it is not to have a mind as well. Early in life, they guess that things without minds might have mindlike properties of intention and that things they will later learn have minds might not possess such intention." (Wegner, 2002, s. 23).

[71] Mechanistyczny oraz mentalny system poznawczy często się przenikają. Jeśli np. widzimy bejsbolistę odbijającego piłkę, to z jednej strony będziemy go postrzegać jako agenta, który realizuje określony cel (chce odbić piłkę), w tej samej scenie dostrzeżemy równocześnie układ funkcjonujący w sposób mechanistyczny: siła uderzenia wprost determinuje odległość, na jaką poszybuje piłka. Widać zatem, że zrozumienie całej sytuacji wymaga dekompozycji działania na określone aspekty, a następnie złączenia rezultatów obydwu strategii poznawczych.

wiedzę dotyczącą traumatycznych przeżyć, lęków, obaw, itp. Te dodatkowe, ważne źródła informacji mogą wpłynąć na zmianę naszej oceny danego zachowania. Sam rezultat często nie wystarcza, by móc prawidłowo ocenić dane działanie, dlatego tak drobiazgowa bywa analiza stosowana przez wymiar sprawiedliwości, który na ogół próbuje zrekonstruować cały kontekst czynu, a nie jedynie jego końcowy efekt. Wszystko to prowadzi do tego, że cele i działania agentów ludzkich mogą być przewidywane z dużym prawdopodobieństwem.

> *Często możemy nauczyć się odczytywać z wyprzedzeniem, co ludzie myślą o swoich działaniach, a czasami informacja taka jest nam dostępna wprost, tak więc jesteśmy w stanie budować złożone zrozumienie prawdopodobnych działań i celów.*[72]

Interpretacja działań własnych, zdaniem Wegnera, odbywa się w taki sam sposób, jak interpretacja działań innych osób. Główna różnica polega jedynie na tym, że mamy uprzywilejowany dostęp do informacji wykorzystywanych podczas procesu interpretacji („luksus" bezpośredniego dostępu do własnych intencji, pragnień, przekonań i planów).

> *Ludzie mają dostęp do skomplikowanego ekranu prezentującego mentalną deskę rozdzielczą wskazującą informacje odnoszące się do celów ich działań, ponieważ wiele wskazówek odnoszących się do działań pojawia się w ich myślach i słowach. Z tego powodu wewnętrzne mechanizmy sprawstwa mogą być dogłębnie zinterpretowane.*[73]

Istotne jest również to, że plany, pragnienia i przekonania, które mają wpływ na nasze zachowanie, nie muszą być wcale uświadamiane w momencie jego realizacji. Wyjątkiem jest intencja, która musi towarzyszyć działaniu. Pozostałe typy treści mentalnych pełnią funkcję „rusztowania" dla intencji – występują w tle, choć niewątpliwie mają wpływ na poczucie sprawstwa. W tym kontekście istotna okazuje się zasada idealnego agenta, do której, zdaniem Wegnera, aspirują wszyscy świadomi agenci. Zgodnie z tą zasadą, jeśli kogoś postrzegamy jako sprawcę, to zakładamy, że jest to agent posiadający określone cele, których jest świadom i które chce osiągnąć, bo traktuje je jako użyteczne. Jest to oczywiście

[72] "We may often learn from people what they think in advance of their actions, and we occasionally have this information available for ourselves as well, so we can construct elaborate understandings of likely actions and goals" (Wegner, 2002, s. 17).

[73] "People have access to an intricate mental dashboard display of cues regarding their goals because lots of cues to their agency appear in their thoughts and words. For this reason the inner workings of their causal agency can be interpreted in great depth" (Wegner, 2002, s. 17).

pewna konstrukcja, która ma nadać sens zaobserwowanemu zachowaniu[74]. Wywiera ona przemożny wpływ również na nasze intencje, a więc na ostateczny sens nadawany działaniom. Potwierdzeniem tego wpływu jest szereg zjawisk dostosowawczych i modyfikujących intencje, które można również interpretować jako ochronę iluzji polegającej na przekonaniu, iż wszystkie zachowania ludzkich agentów są przez nich chciane i świadomie planowane. Osoba, która nie potrafi powiedzieć, dlaczego coś zrobiła lub robi, postrzegana jest jako nieświadoma, odurzona lub chora.

Świadomość przyczyn własnego działania jest jedną z istotnych charakterystyk agenta, dlatego ludzie poświęcają dużo energii, by dysponować odpowiednim wyjaśnieniem własnego zachowania i często przypisują sobie intencje, których faktycznie nie żywili w chwili podejmowania czynności. Znamienny jest tu przykład osoby podlegającej posthipnotycznej sugestii. Osoba obudzona z hipnozy – pod wpływem polecenia: „kiedy się obudzisz, przełożysz książkę ze stołu na regał" – na ogół posłusznie wykonuje narzucone zadanie, a na pytanie o powód takiego działania wyjaśnia: „nie lubię, kiedy rzeczy nie leżą na swoim miejscu". Mamy tu zatem do czynienia z typowym postfaktycznym utworzeniem intencji tłumaczącej podjęte zachowanie[75]. Wegner uważa, że mechanizm wynajdowania intencji zależy głównie od naszych oczekiwań i od nastawienia. Jeśli w kontekście jakiegoś działania oczekiwalibyśmy określonej intencji, a ona nie wystąpi, wówczas tworzymy ją po pojawieniu się jego rezultatu.

Niewątpliwie dzieje się tak na skutek oddziaływania zasady idealnego agenta, której potwierdzenie można znaleźć w obserwacjach małych dzieci. Na podstawie takich obserwacji zakładamy, że działania dowolne wykształcane są w toku rozwoju ontogenetycznego, choć ich pierwotne formy dostrzega się już u noworodków. Na przykład, w eksperymencie z okularami wytwarzającymi iluzję przedmiotów zazwyczaj widzimy następującą sytuację: gdy dziecko nie może chwycić widzianego, iluzyjnie wytworzonego przedmiotu, to zaczyna płakać. Natomiast, gdy to samo działanie dotyczy przedmiotów rzeczywistych – wówczas nie pojawia się płacz u dziecka. Świadczy to o

[74] Por. "We perceive minds by using the idea of an agent to guide our perception. In the case of human agency, we typically do this by assuming that there is an agent that pursues goals and that the agent is conscious of the goals and will find it useful to achieve them. All this is a fabrication, of course, a way of making sense of behavior." (Wegner, 2002, s. 146).

[75] Warto nadmienić, że takie wyjaśnienie nie zawsze się zdarza. Osoba, która wykonuje działanie pod wpływem sugestii, czasami nie potrafi wyjaśnić swego zachowania. Prawdopodobnie mają na to wpływ okoliczności, które prowokują do wytworzenia intencji lub przeciwnie, poprzez niezwykłość, blokują jej powstanie.

silnym związku występującym w umyśle dziecka między chceniem (intencją) a oczekiwanymi rezultatami działania.

Powyżej przedstawiono jedynie wybrane wątki wnikliwych analiz Wegnera, które dotyczą procesów odpowiedzialnych za rozumienie działań własnych oraz działań innych agentów powiązanych z tworzeniem się poczucia sprawstwa.

Podsumowanie

John Searle podczas jednego z wykładów podzielił się ciekawą obserwacją:

> *Filozofia umysłu to pod wieloma względami szczególnego rodzaju gałąź filozofii. W większości działów filozofii możemy bowiem mówić o współbrzmieniu zdrowego rozsądku oraz tego, co akceptują profesjonaliści. W filozofii umysłu jest inaczej, jak sądzę, występuje tu radykalne zerwanie [między zdrowym rozsądkiem a wiedzą ekspercką – uzupełnienie M.C.]. Większość ludzi akceptuje jakąś wersję dualizmu. Są przekonani, że ich życie składa się z dwóch sfer: mentalnej i fizycznej. Dualizm wydaje się mieć jakiś szczególny urok. Jednak wśród nauk specjalizujących się w tym obszarze, w filozofii umysłu, naukach kognitywnych i psychologii dualizm został niemal powszechnie odrzucony.* (Searle, 2011).

Spostrzeżenie Searle'a w dużym stopniu odnosi się również do naszych działań. Zdroworozsądkowy obraz dotyczący ludzkiej woli czy świadomej kontroli zachowań w istotny sposób odbiega od obrazu, który wyłania się z naukowych badań nad prostymi działaniami intencjonalnymi. Zaprezentowane powyżej wyniki wyraźnie odsłaniają złożoną strukturę nawet bardzo prostych aktów wolicjonalnych. Poszczególne składowe intencji w działaniu oraz towarzyszące działaniom poczucie sprawstwa – to odrębne fenomeny, których rozpoznanie w dużym stopniu wymyka się „nieuzbrojonej" w metody eksperymentalne introspekcji. W tej sytuacji nie dziwi fakt, iż argumentacja samego Searle'a na rzecz przyczynowego statusu intencji w działaniu jest niezgodna z wynikami uzyskanymi w naukach empirycznych. Dane eksperymentalne wskazują jedynie na korelacyjny status zależności intencja w działaniu - zachowanie, a nie na jej przyczynowy charakter, jak utrzymuje Searle (omówiłem tę kwestię wyżej w punkcie „Intencja w działaniu jako korelat procesów przygotowawczych").

Wątpliwości budzi również postulowana przez niego struktura warunków spełniania definiująca treść intencji w działaniu. Badania psychologów intencji pokazują, że warunki spełniania wskazane przez Searle'a (reprezentacja zachowania oraz świadomość, że intencja jest przyczyną zachowania) – to w istocie dwa niezależne zjawiska: pierwsze to tzw. chęć wykonania ruchu (składowa motoryczna), a drugie to poczucie sprawstwa, które konstruowane jest albo w trybie predykcyjnym (Haggard), albo rekonstrukcyjnym (Wegner).

Pomimo ewidentnego postępu, ciągle trudno uznać badania nad prostymi działaniami intencjonalnymi za zakończone sukcesem. Nadal nie mamy pewności, że zidentyfikowane zostały wszystkie istotne składowe decydujące o ich przebiegu, nie wiemy również, jakie dokładnie są związki między nimi. Wątpliwości budzi m.in. czysto „laboratoryjny" scenariusz eksperymentów realizowanych zgodnie z instrukcją Libeta. Nie jest wcale oczywiste, jak odnieść zamierzone i zgodne z instrukcją badacza wykonywanie ruchów palcem lub nadgarstkiem do sytuacji, gdy podobne ruchy wynikają z realnych potrzeb i celów agenta. Jest czymś zadziwiającym, jak zauważa Patrick Haggard, że psychologia intencji rozwinęła się całkowicie niezależnie od psychologii nagrody i motywacji (por. Haggard, 2005). Podobna trudność dotyczy wpływu przekonań na kształt intencji. W koncepcji Daniela Wegnera, która dotyczy rekonstruktywistycznego poczucia sprawstwa, przekonania mają charakter jawny (patrz: zasada spójności, zasada idealnego agenta, procesy związane z mentalnym systemem poznawczym, mechanizmy projekcji), ale pełnią one głównie funkcję narracyjną, dopowiadającą i uspójniającą zaobserwowane efekty, a nie determinującą. Warto w tym kontekście zwrócić również uwagę na inne rozłożenie akcentów. W koncepcji predykcyjnej, sprawstwo służy głównie kontroli zachowań, natomiast w ujęciu rekonstruktywistycznym zostaje ono włączone w szerszy, psychologiczny kontekst, w którym realizowane są określone potrzeby podmiotu (patrz: zasada idealnego agenta oraz związane z nią mechanizmy dostosowawcze).

Wątpliwości budzi również nieprecyzyjny i często synonimiczny charakter wielu pojęć wykorzystywanych przez psychologów intencji. Zwraca na to uwagę filozof, Tim Baynes. W przeprowadzonej przez niego krytycznej analizie *The illusion of conscious will* zarzuca jej autorowi, że nie ma obecnie podstaw, by redukować poszczególne aspekty sprawstwa (*sense of agency*) lub fenomeny pokrewne nazywane: „poczuciem działania" (*the experience of doing*), „poczuciem wysiłku" (*the experience of effort*), „przeżyciem świadomej woli" (*experience of conscious will*), „doświadczeniem poczucia wolnej woli"

(*experience of experience of free will*) do doświadczenia świadomej woli (*the experience of conscious will*) (Bayne, 2006).

Ostatnia kwestia, którą warto rozważyć w kontekście przedstawionych badań, to problem prawomocności generalizacji sformułowanych w ramach psychologii intencji. Dominująca strategia postępowania jest w przybliżeniu następująca: na początku przeprowadzany jest eksperyment dotyczący określonego aspektu prostego działania intencjonalnego; na tej podstawie rozstrzyga się czy uzyskany efekt jest istotny statystycznie; na ostatnim etapie wnioski z uzyskanych pomiarów rozciąga się na dowolne działania, bez różnicowania czy są one proste, czy złożone, np. stwierdza się, że chęć wykonania ruchu (*sense of urge*) zawsze pojawia się z opóźnieniem 330 ms w stosunku do potencjału gotowości. Domyślnie zakłada się, że działanie złożone to nic innego, jak sekwencja działań prostych. Taka strategia wyjaśniająca oparta jest na wątłych podstawach. Z badań nad złożonymi strukturami wiemy, że niezwykle rzadko zdarza się tak, by ich zachowania były prostą kombinacją zachowań prostych. Na ogół w tego typu systemach lepiej sprawdza się podejście odwrotne, uznające zachowania proste za radykalnie uproszczone przypadki pewnego ogólnego mechanizmu zaprojektowanego dla przypadków złożonych.

Z niektórymi z wymienionych tu trudności zmierzę się w ostatnim rozdziale, w którym przedstawię zintegrowany model działań intencjonalnych. Koncepcja sprawstwa Wegnera wraz z korelacyjnym statusem intencji w działaniu pełnią w tym modelu zupełnie nową, niezwykle istotną rolę, mianowicie – wspomagają proces przekształcania określonych układów reprezentacji elementarnych w stany intencjonalne. W ten sposób, zrozumiałe stanie się, dlaczego początkowo niemal pusta sieć stanów intencjonalnych zaczyna się stopniowo rozwijać. Okaże się też, że z czasem jej zasoby powiększają się do tego stopnia, iż pozwalają agentowi na deliberację i tworzenie planów. Dokładniejszy opis, na czym polega reprezentacjotwórczy status sprawstwa, będzie jednym z ważniejszych zagadnień podjętych w ostatnim rozdziale pracy.

5 Zintegrowany model złożonego działania intencjonalnego

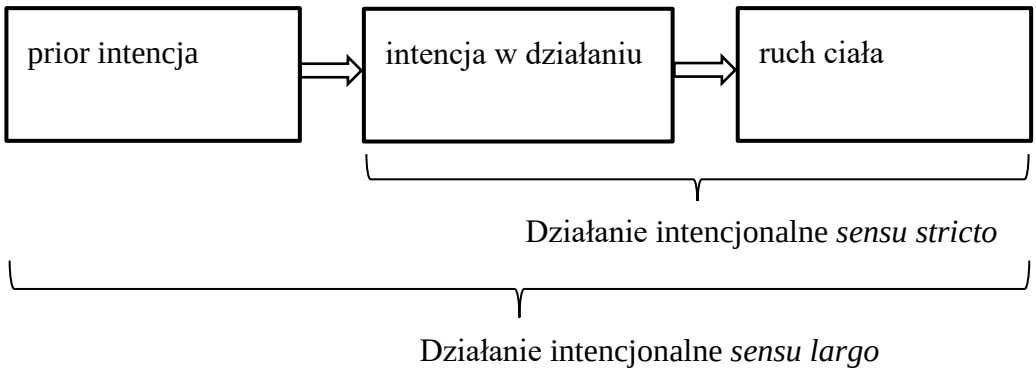
W niniejszym rozdziale przedstawię i omówię model złożonego działania intencjonalnego. Zanim uzasadnię, dlaczego złożone działania intencjonalne wybrałem jako obiekt modelowania, przywołam stanowisko naturalizmu biologicznego Searle'a (por. rozdział 2. niniejszej pracy) oraz przeanalizuję jego przydatność do charakterystyki działania intencjonalnego. Następnie przypomnę Searle'owskie rozróżnienie na proste i złożone działanie intencjonalne i wyjaśnię, dlaczego uważam, że potrzebne jest alternatywne ujęcie tej dystynkcji.

5.1 Problem naturalizacji umysłowych składników działania intencjonalnego

Na wstępie przypomnę najważniejsze idee Searle'owskiej koncepcji działania intencjonalnego. Pozwoli to pokazać, które z jego idei są przydatne (raczej jako inspiracja, niż jako tezy dające się bezpośrednio wykorzystać) przy budowie modelu działania intencjonalnego, a których – ze względu na ich spekulatywność albo mglistość – nie uda się wykorzystać. Za szczególnie inspirujące uważam jego rozbudowane ujęcie struktury działania intencjonalnego. Searle podjął się powiązania filozoficznego pojęcia intencjonalności z pojęciem zachowania biologicznego. To pierwsze odniósł do zdecydowanej większości stanów umysłowych, to drugie – do fizycznego ruchu organizmu. Searle, wiążąc te dwa pojęcia, charakteryzuje działanie intencjonalne jako złożone z zamiaru (intencji) oraz z zachowania. Intencjonalność zamiaru polega na tym, że jest on stanem umysłowym skierowanym na określony obiekt lub na stan rzeczy. Określone zachowanie, czyli fizyczny ruch organizmu ma zapewnić osiągnięcie intencjonalnie zamierzonego celu. Searle postuluje, aby dookreślić pojęcie „zamiaru" poprzez

odróżnienie stanu umysłowego wyprzedzającego zachowanie od stanu, który towarzyszy zachowaniu. Ten pierwszy typ zamiaru nazywa on prior intencją, ten drugi – intencją w działaniu. Prior intencja jest zgrubnym projektem działania, które doprowadzić ma do wystąpienia pożądanego stanu rzeczy. Natomiast intencja w działaniu to stan umysłowy nierozerwalnie zespolony z samym zachowaniem. Jej zadaniem jest wyznaczanie „na bieżąco" sposobu, za pomocą którego agent będzie realizował zachowanie. Pojęcie prior intencji odpowiada tradycyjnemu rozumieniu intencji jako stanu umysłowego projektującego działanie, które może być podjęte w bliższej lub dalszej przyszłości. Natomiast istotnym *novum* propozycji Searle'a jest pojęcie intencji w działaniu. Ten typ stanu umysłowego pojawia się równocześnie z samym zachowaniem, a jego istotą jest odczuwanie działania (*experience of acting*), a zarazem – za sprawą informacji pozyskiwanych w trakcie odczuwania – „sterowanie" jego przebiegiem do zamierzonego, a więc wyznaczonego przez intencję w działaniu, ruchu kończącego jej realizację. Zdaniem Searle'a, każde działanie intencjonalne składa się z intencji w działaniu oraz z fizycznego ruchu, natomiast nie wszystkie z działań intencjonalnych są poprzedzone prior intencją. Można zatem powiedzieć, iż odróżnia on działanie intencjonalne w wąskim oraz szerokim sensie.[76] To pierwsze, czyli działanie intencjonalne *sensu stricto*, to para złożona z intencji w działaniu oraz z zachowania. To drugie, czyli działanie intencjonalne *sensu largo* – to para, której pierwszym członem jest prior intencja, a drugim – działanie intencjonalne *sensu stricto*.

Poniższy diagram obrazuje zaproponowane przez Searle'a ujęcie struktury działania intencjonalnego:



[76] W pracach Searle'a dystynkcja między wąskim a szerokim pojmowaniem działania intencjonalnego nie została wprowadzona *explicite*. Jednak zarówno argumentacja, jak i podawane przez niego przykłady pozwalają przyjąć, że faktycznie respektował takie rozróżnienie.

Diagram 9. Główne składowe działania intencjonalnego wg. Searle'a.

W rozdziale 2. wskazałem kolejne składniki, które są potrzebne Searle'owi do rozbudowy struktury koncepcji działania intencjonalnego w szerokim sensie: deliberację, sieć stanów intencjonalnych oraz tło. Nie ulega wątpliwości, że są to ważne komponenty kontekstu, w którym zanurzone jest działanie intencjonalne. Searle nie pokazuje jednak dokładniej, w jaki sposób te intencjonalne (deliberacja, sieć stanów intencjonalnych), a także nieintencjonalne (tło) składowe kontekstu wpływają na realizację całego działania.

Przedstawiona wyżej charakterystyka działania intencjonalnego zdaje się wskazywać, że jej autor jest dualistą uznającym niezależne istnienie stanów umysłowych oraz fizycznych. Sam Searle odrzuca jednak taką supozycję i – aby nie zostać uznanym za reprezentanta któregoś z dotychczasowych stanowisk wyróżnianych w filozofii umysłu (monizm, monizm anomalny, dualizm własności, itp.) – postuluje nowe ujęcie, które nazywa „naturalizmem biologicznym" (Searle, 1983 oraz 2007). Kluczowa w tym ujęciu okazuje się następująca teza: intencjonalność, rozumiana powszechnie przez filozofów jako konstytutywna cecha tego, co umysłowe, a więc niefizyczne, jest w istocie własnością fizyczną, a dokładniej – własnością szczególnej aktywności mózgu. Searle uważa, że ten sam organ, którym jest mózg, może być opisywany na różnych poziomach, zarówno na niskim poziomie wyładowań neuronów i wydzielania neuroprzekaźników, jak i na wysokim poziomie stanów umysłowych. Te ostatnie są również stanami mózgu tyle, że powstawać mają na skutek wysokopoziomowych interakcji między strukturami tego najbardziej złożonego narządu. Tak rozumiany naturalizm biologiczny ma być, w opinii Searle'a (2007, s. 325), odpowiedzią na filozoficzny problem relacji umysł-ciało. Tę, biologicznie rozumianą, intencjonalność uczony odnosi także do działania intencjonalnego:

> *Zazwyczaj, kiedy np. podejmuję świadomą decyzję, aby podnieść rękę i moja ręka unosi się w górę, to moja decyzja wywołuje ruch mojej ręki ku górze. Tak jak każdy system fizyczny, także mózg daje się opisać na różnych poziomach. Wszystkie one to realne przyczynowo poziomy jednego i tego samego systemu przyczynowego. Możemy zatem opisać unoszenie się mojego ramienia na poziomie świadomej intencji-w-działaniu, nakierowanej na podniesienie ręki wraz z odpowiadającym jej ruchem ciała. Możemy też opisać to na poziomie wyładowań neuronów i synaps oraz wydzielania acetylocholiny na płytki końcowe aksonów moich neuronów ruchowych.*

> *Podobnie, możemy opisać pracę silnika samochodu: na poziomie tłoków, cylindrów i wyładowań świec zapłonowych, ale także możemy opisać ją na poziomie utleniania cząsteczek węglowodorów i molekularnej struktury stopów metali. Zarówno w przypadku opisów mózgu, jak i silnika samochodu nie są to odrębne struktury przyczynowe; jest to jedna struktura przyczynowa opisana na różnych poziomach. Kiedy zrozumiesz, że ten sam system może mieć różne poziomy opisu, które nie konkurują ze sobą ani nie są odrębne, ale są tylko różnymi poziomami w obrębie jednego zunifikowanego systemu przyczynowego, to fakt, że mózg ma różne poziomy opisu, nie jest bardziej tajemniczy niż to, że każdy inny system fizyczny ma różne poziomy opisu.*[77]

Argumentacja Searle'a wydaje się nie do odparcia. Wszak jest oczywiste, że podłożem procesów umysłowych są procesy mózgowe, dlaczego więc nie przyjąć, że te pierwsze są szczególnego rodzaju procesami mózgowymi? Takiego zdania był np. Francis Crick, który sformułował to dobitnie i nazwał: „zdumiewającą hipotezą":

> *Zdumiewająca hipoteza brzmi: Ty, Twoje radości i smutki, Twoje wspomnienia i ambicje, Twoje poczucie tożsamości i wolna wola, nie są w rzeczywistości niczym innym niż sposobem, w jaki zachowuje się ogromny zbiór komórek nerwowych i związanych z nimi cząsteczek*. (Crick 1997, s. 17).

Hipoteza Cricka[78] wydaje się być zbieżna, jeśli nie tożsama z naturalizmem biologicznym Searle'a. Sam twórca tej nowej odmiany naturalizmu uznawał neurobiologów takich jak Crick, Koch czy Edelman za zwolenników jego propozycji.[79] Uważam, że podobieństwa

[77] „Typically, for example, when I make a conscious decision to raise my arm and my arm goes up, my decision causes my arm to go up. As with all physical systems, the brain admits of different levels of description, all of which are causally real levels of one and the same causal system. Thus we can describe my arm going up at the level of the conscious intention- in-action to raise my arm, and the corresponding bodily movement, or we can describe it at the level of neuron firings and synapses and the secretion of acetylcholine at the axon endplates of my motor neurons, just as we can describe the operation of the car engine at the level of pistons, cylinders, and spark plugs firing, or we can describe it at the level of the oxidization of hydrocarbon molecules and the molecular structure of metal alloys. In both the case of the brain and the case of the car engine, these are not separate causal structures; it is a single causal structure described at different levels. Once you see that the same system can have different levels of description which are not competing or distinct, but rather different levels within a single unified causal system, the fact that the brain has different levels of description is no more mysterious than that any other physical system has different levels of description." (Searle, 2007, s. 328).

[78] Crick przyznawał, że nie jest autorem tego poglądu i wskazywał artykuł Horace'a Barlowa z 1973 roku, w którym jasno wyrażono taką właśnie ideę (Crick, 197, s. 22).

[79] "It is worth pointing out that practicing neurobiologists of my acquaintance, such as the late Francis Crick, Gerald Edelman, and Christof Koch, implicitly or explicitly accept a version of what I have been calling Biological Naturalism." (Searle 2007, s. 334) „Warto podkreślić, że czynni neurobiolodzy, z jakimi

między poglądami Cricka i Searle'a są pozorne. Co prawda, Crick twierdzi, że ludzkie stany umysłowe można sprowadzić do „zachowania się ogromnego zbioru komórek nerwowych", lecz nie zamierza on za pomocą tego stwierdzenia proponować rozwiązania problemu umysł-ciało. Jego zdumiewająca hipoteza jest wezwaniem do budowania programu badawczego, którego celem będzie znalezienie korelatów neuronalnych świadomych stanów umysłowych.[80] Innymi słowy, Crick uważa, że poszukiwanie wyjaśnienia dla stanów umysłowych wymaga naukowej eksploracji mózgu, bo tylko w ten sposób procesy umysłowe zakotwiczone zostaną w procesach biologicznych. Co więcej, nie jest to tylko głoszony przez niego manifest, ale odwołanie do własnej praktyki badawczej, którą autor *Zdumiewającej hipotezy* traktował jako naukowe poszukiwanie neuronalnych korelatów świadomości. Zauważmy, że takie podejście radykalnie odbiega od sposobu postępowania Searle'a. Twórca koncepcji naturalizmu biologicznego dostarcza filozoficznej odpowiedzi na pytanie o status takich cech stanów umysłowych, jak intencjonalność czy świadomość. Co prawda, twierdzi on, że cechy te, jak i wyposażone w nie stany umysłowe są stanami biologicznymi, występującymi na wysokich poziomach działania mózgu, jednak nie pokazuje, jak powiązać charakteryzowane filozoficznie stany umysłowe z dającymi się opisać w języku biologii, wysokopoziomowymi stanami mózgu. Nie wystarczy nazwać stanów intencjonalnych stanami mózgu, aby ogłosić, że posiadające cechę intencjonalności stany umysłowe zostały znaturalizowane. Należałoby jeszcze objaśnić, na czym naturalizacja taka miałaby polegać.

Od filozofa proponującego zakotwiczenie intencjonalności w procesach biologicznych należałoby oczekiwać, że doprecyzuje, jak można wykorzystać wiedzę biologiczną do wyjaśnienia procesów umysłowych. Innymi słowy, autor tezy o biologicznej naturze intencjonalności nie wyjaśnia, jak stany czy procesy umysłowe powiązać z rozumianymi na sposób biologiczny stanami czy procesami występującymi na wysokich poziomach organizacji mózgu. „Delegowanie" na samych naukowców, by w języku ich dziedziny dookreślili oni, o jakie stany mózgu w tym przypadku chodzi, świadczy o tym, iż filozof podkreślający silne związki swojej koncepcji z wiedzą z nauk biologicznych poprzestaje jedynie na poziomie intuicji badawczych.

się zetknąłem, tacy jak niedawno zmarły Francis Crick, a także Gerald Edelman czy Christof Koch, *implicite* lub *explicite* akceptują jakąś wersję tego, co nazywam biologicznym naturalizmem."

[80] Przedmiotem szczególnego zainteresowania badawczego Cricka były stany świadome, a dokładniej – świadomość wzrokowa.

Wyróżnić można trzy słabości dyskutowanego tu naturalizmu biologicznego:

(a) Spekulatywny charakter tej koncepcji.
(b) Uznanie opisu zjawiska za jego wyjaśnienie.
(c) Nietrafna charakterystyka wielopoziomowych podejść badawczych.

Szczegółowa dyskusja wymienionych niedostatków naturalizmu biologicznego Searle'a wykracza poza tematykę niniejszej rozprawy. Omówię je w takim stopniu, w jakim jest to niezbędne dla pokazania, że znaturalizowana charakterystyka działania intencjonalnego wymaga wykroczenia poza naturalizm biologiczny i podjęcia pracy od tego miejsca, które dla Searle'a było jej zakończeniem.

(ad a)

Naturalizm biologiczny zdaje się być solidnie osadzony w neurobiologicznej wiedzy o strukturze i funkcjach mózgu. Jednak bliższa analiza argumentacji na rzecz tego stanowiska przekonuje, że jego autor w gruncie rzeczy poprzestaje na sformułowaniu ogólnikowej tezy, iż stany umysłowe to cechy przysługujące strukturom lub funkcjom występującym na wyższych poziomach organizacji mózgu. Ustosunkowanie się do takiej tezy wymagałoby jej dookreślenia. W pracach Searle'a próżno szukać jej doprecyzowania. Natomiast przytaczane przez niego przykłady, które mają uargumentować tę tezę, są nader uproszczone i odwołują się do podręcznikowej wiedzy o stanach mózgu determinujących stany umysłowe.[81] Z tego też powodu uważam, iż stanowisko naturalizmu biologicznego jest – wbrew nazwie i oświadczeniom jego twórcy – wytworem filozofii spekulatywnej. O tym, jak Searle pojmuje przydatność ustaleń naukowych w badaniach (świadomych) stanów umysłowych – świadczy choćby ta jego wypowiedź:

> *Nie znamy wszystkich szczegółów, w jaki dokładnie sposób procesy mózgowe wywołują świadomość, ale nie ma wątpliwości, że tak jest. Za tezą, że wszystkie nasze*

[81] Warto dodać, że spekulatywne podejście do tematyki percepcji intensywnie rozwijanej we współczesnej nauce cechuje także późniejsze rozważania Searle'a. Josh Armstrong w swojej recenzji Searle'a teorii percepcji *Seeing Things as They Are: A Theory of Perception* (Searle, 2015) tak oto charakteryzuje jego podejście: „Searle develops his theory of perception from what we call the armchair, by which I mean that Searle develops the details of his positive account in almost complete isolation from the wealth of recent work in perceptual psychology. If perception is, as Searle insists, a natural kind like digestion or photosynthesis, then one cannot provide a theory of its operations or attempt to answer philosophical questions about its nature independently from empirical investigation. In short, Searle cannot have it both ways: he must either give up his naturalism or radically revise what he takes to constitute a theory of the relevant domain." (Armstrong, 2015).

> *świadome stany – od odczuwania pragnienia do przeżywania mistycznych ekstaz – są wywołane przez procesy mózgowe, przemawia przytłaczająca liczba danych empirycznych. W istocie najbardziej obecnie ekscytujące badania w naukach biologicznych – to próby ustalenia, jak to się dokładnie dzieje. Jakie są neuronalne korelaty świadomości i jak działają, aby wywoływać stany świadome?* [82]

Powyższy cytat pokazuje, że Searle faktycznie formułuje swój naturalizm biologiczny w sposób ogólnikowy, a zadanie jego doprecyzowania pozostawia przyszłym badaniom neuronaukowym. Tak zarysowany podział pracy ma, jak się wydaje, swoje korzenie w bardzo krytycznym stosunku Searle'a do ustaleń nowożytnej filozofii umysłu. W jego opinii, wiele z rozstrzygnięć filozoficznych nawiązuje do kartezjańskiego rozróżnienia na to, co materialne i na to, co mentalne, przez co wikła badania naukowe w spory pojęciowe, które nie są adekwatne do współczesnej wiedzy neurobiologicznej.

W niniejszym rozdziale opowiadam się za nieco odmiennym pojmowaniem naturalistycznego podejścia do relacji: filozofia – rozstrzygnięcia nauk empirycznych. Do jego istoty należy wykorzystanie wiedzy naukowej, aby za jej pomocą przekształcić wyjściowe intuicje filozoficzne w tezy (modele), które poddają się naukowej krytyce i analizie.

Dwa pozostałe niedostatki koncepcji naturalizmu biologicznego są również efektem skupienia się na filozoficznej charakterystyce stanów umysłowych i braku podjęcia bardziej wnikliwych prób pokazania, w jakiej relacji stany te pozostają do faktycznych struktur mózgu i prawidłowości ich działania.[83]

(ad b)

[82] "We do not know all the details of exactly how consciousness is caused by brain processes, but there is no doubt that it is in fact. The thesis that all of our conscious states, from feeling thirsty to experiencing mystical ecstasies, are caused by brain processes is now established by an overwhelming amount of evidence. Indeed, the currently most exciting research in the biological sciences is to try to figure out exactly how it works. What are the neuronal correlates of consciousness and how do they function to cause conscious states?" (Searle 2007, s. 328)

[83] Brak troski o ugruntowanie naturalizmu biologicznego we współczesnej wiedzy naukowej widać wyraźnie w przywoływanym tu artykule „Biological Naturalism" (Searle 2007). Dołączona do niego literatura liczy pięć prac, wszystkie zostały napisane przez autora artykułu. Co prawda, w samym tekście przywołano nazwiska trzech uznanych badaczy (Crick, Koch, Edelman), ale tylko po to, aby poinformować, że – według Searle'a – są oni zwolennikami prezentowanego w tekście naturalizmu biologicznego (Searle 2007, s. 334).

Najczęściej przywoływanym przykładem, za pomocą którego Searle obrazuje ideę naturalizmu biologicznego, jest intencjonalne podniesienie ręki. Uczony, omawiając ten przykład, stwierdza, że unoszenie ręki ku górze może być opisane dwojako: „na poziomie wyładowań neuronów i synaps oraz wydzielania acetylocholiny na płytki końcowe aksonów moich neuronów ruchowych" oraz „na poziomie świadomej intencji-w-działaniu, nakierowanej na podniesienie ręki wraz z odpowiadającym jej ruchem ciała" (Searle 2007, 328). Z dalszego wywodu dowiadujemy się, że przedmiotem tych dwóch opisów jest jedna i ta sama struktura przyczynowa, jaką jest mózg, tyle – że każdy z opisów charakteryzuje tylko te składniki i procesy, które znajdują się na jednym, wybranym poziomie organizacji całego systemu. Pomińmy tu kwestię wielości poziomów, wrócę do niej za chwilę, i zwróćmy uwagę, na to, że w przywołanym przykładzie mowa jest o różnych opisach jednego obiektu. Trudno sobie wyobrazić, aby opis działania intencjonalnego za pomocą pojęć filozoficznych (takich jak prior intencja czy intencja w działaniu) mógł zostać odniesiony bezpośrednio do charakterystyki mózgu, rozumianego jako narząd organizmu ludzkiego badany przez nauki biologiczne. Frazy „mózg powziął zamiar podniesienia ręki" (prior intencja) czy „mózg zamierza kontynuować podnoszenie ręki" nie tylko brzmią dziwacznie, ale rozumiane literalnie znaczyłyby, że równocześnie z pojawieniem się stanu umysłowego o charakterze zamiaru (albo tuż po jego wystąpieniu) pojawia się stosowna zmiana w aktywności mózgu, która jest przyczynowo powiązana z podniesieniem ręki. Obraz taki jest zgodny z ujęciem prezentowanym przez Searle'a, szkopuł w tym, że nie jest zgodny z wynikami badań empirycznych. Mam tu przede wszystkim na względzie eksperyment Libeta, zrelacjonowany w rozdziale czwartym niniejszej pracy. Przypomnijmy, że w eksperymencie tym badani sami swobodnie decydowali o tym, w którym momencie będą zginali palce. Jednak moment podjęcia decyzji o rozpoczęciu ruchu bynajmniej nie był tożsamy z pojawieniem się aktywności w korze motorycznej, która to aktywność wskazuje na uruchomienie programu motorycznego odpowiedzialnego za zgięcie palca. Aktywność mózgu – nazywana potencjałem gotowości – wyprzedzała świadomy zamiar o co najmniej pół sekundy. Niezależnie od rozmaitych krytycznych interpretacji wyniku Libeta jedno jest pewne: o zgięciu palca nie decydował świadomy zamiar, ale poprzedzający go stan gotowości mózgu. Zatem przyczyną ruchu palca nie mogła być intencja w działaniu (w momencie powzięcia zamiaru nie zarejestrowano aktywności mózgu), lecz wcześniejsza aktywność „decyzyjna" mózgu. Zgięcie palca, podobnie jak podniesienie ręki, to przykłady prostych – w potocznym rozumieniu tego słowa – działań intencjonalnych. Można zatem odnieść ustalenia eksperymentu Libeta do

podanego przez Searle'a przykładu z podnoszeniem ręki. W efekcie, sformułować można dwa wnioski.

Po pierwsze, skoro to nie decyzja o zgięciu palca jest przyczyną jego ruchu, to – analogicznie – decyzja i towarzysząca jej intencja w działaniu, pojmowana w sposób Searle'owski, nie mogą być przyczyną ruchu ramienia ku górze.

Po drugie, opis aktywności mózgu, jaką jest tworzenie się potencjału gotowości, choć ciągle niedoskonały, w żaden sposób nie przybliża nas do zrozumienia roli, jaką odgrywa intencja w działaniu. Błąd Searle'a polega na tym, iż traktuje opis aktywności mózgu jako wystarczające wyjaśnienie stanów intencjonalnych. Zdaje się nie dostrzegać, że stany czy procesy mózgowe, które można opisać w języku neuronauki, same podlegają wyjaśnieniu. Takie wyjaśnienie wymaga przynajmniej wskazania mechanizmu, który wywołuje takie, a nie inne zmiany w procesach mózgowych. Dopiero rekonstrukcja odpowiednich mechanizmów decydujących o sposobie działania systemów mózgowych, a nie jedynie opisanie przejawów ich działania, może pomóc w wyjaśnieniu funkcji, które faktycznie pełnią stany intencjonalne. Nie wystarczy opisać mózgowych korelatów intencjonalnych stanów umysłowych, trzeba jeszcze określić, jak ich organizacja wpływa na ludzkie życie umysłowe. Naturalizm biologiczny w jego oryginalnym sformułowaniu ogranicza się wyłącznie do postulowania opisu stanów mózgu, pomija natomiast kwestie funkcji, które pełnią jego podsystemy.

Zauważmy, że funkcje podsystemów mózgu nie dają się łatwo scharakteryzować w języku czysto biologicznym. Buduje się wyidealizowane modele, które często mają charakter obliczeniowy, by uchwycić istotę działania takiego podsystemu. Pożądane jest, aby obliczeniowe modele miały swoje implementacje neuronalne. Takie właśnie podejście: budowanie obliczeniowych modeli i wskazywanie – tam, gdzie to możliwe – ich implementacji neuronalnych leży u podstaw proponowanego tu modelu złożonego działania intencjonalnego.

(ad c)

Ostatnia z uwag o słabościach naturalizmu biologicznego Searle'a odnosi się do jego stwierdzenia, że ten sam system może być opisany na kilku różnych poziomach:

> *Kiedy zrozumiesz, że ten sam system może mieć różne poziomy opisu, które nie konkurują ze sobą, ani nie są odrębne, ale są tylko różnymi poziomami w obrębie*

> *jednego zunifikowanego systemu przyczynowego, to fakt, że mózg ma różne poziomy opisu, nie jest bardziej tajemniczy niż to, że każdy inny system fizyczny ma różne poziomy opisu*. (Searle 2007, s. 328).

Literalna interpretacja tej wypowiedzi skłania do uznania jej za mało odkrywczą. Trudno znaleźć przeciwników poglądu, że struktury systemów złożonych, a mózg – jak wiadomo – jest niezwykle złożonym systemem, są wielopoziomowe. Oczywiste jest również to, że każdy z tych poziomów może być przedmiotem odrębnego opisu, i – co więcej – opis podsystemów z poziomu niższego pokazuje – z jakich składników złożony jest podsystem nadrzędny wobec nich. Jeśli jednak wypowiedź tę usytuować w kontekście wcześniejszych uwag Searle'a, to przyjąć można, że chce on przekazać wysoce oryginalną myśl, a mianowicie: na jednym z poziomów mózgu występują stany umysłowe. Problem polega na tym, że jej autor nie precyzuje ani tego, jak rozumie pojęcie poziomu[84], ani – jakie to systemy biologiczne (pamiętajmy, że chodzi tu o naturalizm biologiczny!) w obrębie narządu, którym jest mózg, miałyby wytwarzać świadome intencjonalne stany umysłowe. Bez takiego, choćby zgrubnego, określenia, do jakiego rodzaju poziomów organizacji mózgu Searle się odnosi, można uznać, że jego propozycja niczym istotnym nie różni się od emergentyzmu (stanowiska, zgodnie z którym w złożonym systemie można wyróżnić jednostki niższego rzędu oraz wynikające z ich kompozycji jednostki wyższego rzędu, a ponadto, że własności jednostek wyższego rzędu (emergentne) nie dają się w prosty sposób wyjaśnić poprzez własności jednostek niższego rzędu (O'Connor, 2020)). Przykład przytaczany przez Searle'a potwierdza wskazane odniesienie. W jego opinii, opis mózgu-umysłu strukturalnie niczym się nie różni od opisu silnika samochodowego, który na elementarnym poziomie jest zbiorem molekuł o określonych własnościach, a na bardziej złożonym poziomie, wynikającym z ich kompozycji, zbiorem części (tłoków, cylindrów, itp.) o specyficznym kształcie, trwałości, itd. Osiągnięcia współczesnej fizyki wskazują, że oba opisy są spójne i można między nimi przeprowadzić systematyczne mapowanie, tzn. określone składowe jednego poziomu można odnieść do składowych drugiego poziomu oraz pokazać, jak te drugie zależą od pierwszych. Co istotne, podobne mapowanie można przeprowadzić w odniesieniu do związków przyczynowych funkcjonujących między

[84] Badaczem, który pokazał, że pojęcie poziomu mózgu można rozumieć przynajmniej na trzy sposoby, jako: (a) poziom analizy, (b) poziom strukturalnej organizacji oraz (c) funkcji, pojmowanej jako przetwarzanie informacji, był David Marr (1982). Odróżnienie między poziomami analizy, organizacji oraz przetwarzania omówione jest w książce Patricii S. Churchland i Terrence'a J. Sejnowskiego (1992) *The Computational Brain*.

składowymi występującymi na danym poziomie. Innymi słowy, możliwa jest pełna redukcja wyższego poziomu do niższego. Analogiczna sytuacja, twierdzi Searle, występuje między poziomem neuronalnym i mentalnym – złożone struktury neuronalne wytwarzają wysokopoziomowe stany mentalne o specyficznych własnościach. Jedyna różnica polega na tym, że w przypadku stanów umysłowych redukcja przyczynowa nie powoduje pełnej redefinicji przedmiotu, który jej podlega (nie dochodzi do tzw. redukcji ontologicznej). Wyjaśnienie stanu umysłowego w kategoriach neuronalnych powoduje, że „traci" on swój jakościowy i subiektywny charakter, np. ból staje się sekwencją wyładowań komórek nerwowych. Przedstawione ujęcie, choć w wielu miejscach trafne, ciągle wydaje się niepełne. Searle jakby nie dostrzega, że samo zredukowanie stanów umysłowych do stanów mózgu nie dostarcza nam nowej, satysfakcjonującej wiedzy o umyśle. Proponowane przez niego rozwiązanie ma charakter fenomenalistyczny i nie przybliża nas do wyjaśnienia tego, jak umysł działa, czyli do odkrycia mechanizmów jego funkcjonowania. W poszukiwaniu wyjaśnień badacze często opuszczają teren biologii i wykorzystują wiedzę z dyscyplin z nią niespokrewnionych. Za taką teoretyzującą postawą opowiada się David Marr.

> *Próba zrozumienia percepcji na podstawie badania dotyczącego samych neuronów jest jak próba zrozumienia lotu ptaków na podstawie badania samych piór. Po prostu nie da się tego zrobić. Aby zrozumieć lot ptaków, musimy zrozumieć aerodynamikę; dopiero wtedy budowa piór i różne kształty skrzydeł ptaków nabierają sensu. Co więcej, jak zobaczymy, nie możemy zrozumieć, dlaczego komórki zwojowe siatkówki i neurony ciała kolankowatego bocznego mają takie pola recepcyjne, jakie mają, badając jedynie ich anatomię i fizjologię. Możemy zrozumieć takie zachowanie tych komórek i neuronów, badając ich obwody i interakcje, ale – by pojąć, dlaczego pola recepcyjne są takie, jakie są - dlaczego są koliście symetryczne i dlaczego ich obszary pobudzające i hamujące mają charakterystyczne kształty i rozkłady - musimy poznać choć trochę teorię operatorów różniczkowych, filtry środkowoprzepustowe i matematykę zasady nieoznaczoności.*[85]

[85] "In a similar vein, trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers. It just cannot be done. In order to understand bird flight, we have to understand aerodynamics; only then do the structure of feathers and the different shapes of birds' wings make sense. More to the point, as we shall see, we cannot understand why retinal ganglion cells and lateral geniculate neurons have the receptive fields they do just by studying their anatomy and physiology. We can understand how these cells and neurons behave as they do by studying their wiring and interactions, but in order to understand why the receptive fields are as they are – why they are circularly symmetrical and why their excitatory and inhibitory regions have characteristic shapes and distributions – we have to know a little

Marr twierdzi, iż zrozumienie, czyli wyjaśnienie zjawisk mózgowych, wymaga wykroczenia poza opis, a wtedy, kiedy poszukiwane są mechanizmy, czyli prawidłowości decydujące o występowaniu tych zjawisk, sięgnąć trzeba po wiedzę spoza biologii. Niestety, naturalizm biologiczny zdaje się nie dostrzegać tego problemu.

Podsumowując, wskazałem i omówiłem trzy słabości naturalizmu biologicznego: (a) deklarowanie potrzeby osadzenia działań intencjonalnych w wiedzy z nauk biologicznych bez zademonstrowania lub choćby zasugerowania, jak należałoby to zrobić, (b) uznanie, że istotą naukowego ujęcia działania intencjonalnego jest jego opis, a nie wyjaśnienie mechanizmów decydujących o jego przebiegu, (c) brak uwzględnienia tego, że jednym z podstawowych zadań przyrodoznawcy badającego obiekty z określonej dziedziny (w tym przypadku obiektami są mózgi) jest wyznaczenie takiego poziomu ich opisu, który pozwala odkryć i scharakteryzować prawidłowości ich funkcjonowania oraz wyjaśnić ich zachowania. Ulokowanie stanów umysłowych na wyższych poziomach organizacji mózgu, bez choćby szkicowego określenia cech takiego poziomu – najistotniejszego dla badanych zjawisk – jest próbą uniknięcia problemu, a nie jego rozwiązaniem.

Omówienie słabości naturalizmu biologicznego pomaga w ich uniknięciu przy konstrukcji modelu działania intencjonalnego. Dlatego też w niniejszym rozdziale zachowana została istota podejścia naturalistycznego, jednak – w odróżnieniu od naturalizmu biologicznego Searle'a – uznałem, iż trzeba, przybliżając za pośrednictwem kolejnych modeli strukturę oraz mechanizmy funkcjonowania działania intencjonalnego, wykorzystać wiedzę spoza nauk biologicznych.

Zanim przystąpię do prezentacji sekwencji modeli, dookreślę pojęcie intencji (zamiaru) i jej przedmiotu intencjonalnego (dla Searle'a był to w ostatniej instancji ruch ciała) oraz uzasadnię, dlaczego niniejsza propozycja skupia się na charakterystyce złożonych działań intencjonalnych.

Rola wartości i uczenia się w złożonym działaniu intencjonalnym

Zamiar i jego odniesienie przedmiotowe

of the theory of differential operators, band-pass channels, and the mathematics of the uncertainty principle." (Marr, 2010, s. 28)

W omówionej w rozdziale 2. i przyjętej tu jako punkt wyjścia, lecz nie dojścia, koncepcji Johna Searle'a odróżnia się działanie intencjonalne w sensie wąskim oraz szerokim. To pierwsze określane jest jako wykonywane aktualnie działanie, na które składają się: stan umysłowy, czyli intencja w działaniu oraz cielesne zachowanie. Jeśli tak pojmowane działanie poprzedzone jest przez powzięty wcześniej zamiar, nazwany prior intencją, to powstały przez jej dołączenie układ jest działaniem intencjonalnym w sensie szerokim. Prior intencja oraz intencja w działaniu są stanami umysłowymi, które cechuje intencjonalność. Własność ta przysługuje większości – pojmowanych standardowo – stanów umysłowych. Pojawia się zatem pytanie: czy zamiar cechuje się szczególnego rodzaju, jemu tylko właściwą, intencjonalnością, czy też intencjonalność zamiaru nie różni się istotnie od intencjonalności percepcji, pamięci, pragnienia, przekonania, przypuszczenia, itp.? Searle nie daje w swoich pracach jednoznacznej odpowiedzi na to pytanie. Z jednej strony trzyma się ogólnej charakterystyki intencjonalności jako uniwersalnej cechy stanów umysłowych[86], z drugiej jednak, kiedy omawia intencjonalność, zarówno prior intencji, jak i intencji w działaniu, to wskazuje na dwie, charakterystyczne dla nich cechy: sprawczość oraz cielesny ruch (zachowanie) jako szczególny rodzaj przedmiotu (odniesienia) intencjonalnego.[87] Sprawczość to cecha decydująca o tym, że intencja ma zdolność do przyczynowego wywoływania cielesnego ruchu. Searle podkreśla, że sprawczość, pojmowana jako zdolność do wywoływania zmian w świecie fizycznym, nie przysługuje innym stanom umysłowym łączonym z działaniem intencjonalnym, takim jak pragnienie (*desire*) czy przekonanie (*belief*). Jego pogląd różni się od szeroko akceptowanego stanowiska, zgodnie z którym to pragnienia wraz z przekonaniami są przyczyną ludzkich działań intencjonalnych. Autor *Intentionality* argumentuje, że do tego, aby pragnienie lub przekonanie zostały spełnione, nie jest konieczne jakiekolwiek działanie podmiotu (kwestię tę omawiam dokładniej w rozdziale 2). Oczywiście, aby zrealizować pragnienia na ogół podejmujemy stosowne działania, jednak gdyby stało się tak, że przedmiot pragnienia zaistniałby bez naszej aktywności, to i tak uznalibyśmy je za spełnione. Sytuacja taka nie jest możliwa w przypadku zamiaru. Jeśli zamierzamy pójść do kina, a znaleźlibyśmy się w nim bez aktywności z naszej strony (np. uśpiono by nas, a następnie przeniesiono do kina i wybudzono), zamiar nie zostałby spełniony. Searle

[86] Por. rozdziału 2 niniejszej pracy.

[87] Dokładniej rzecz biorąc, ruch ciała jest przedmiotem intencjonalnym intencji w działaniu, ale pośrednio także prior intencji, gdyż jej przedmiotem intencjonalnym jest działanie intencjonalne sensu stricto, którego składnikiem jest ruch ciała.

argumentuje, że spełnienie zamiaru wymaga wystąpienia ruchu ciała i to w odpowiednich okolicznościach i stosownym czasie. Uczony ilustruje te wymogi przywołując eksperyment Penfielda (Searle, 1980, s. 57). Badacz ten pobudzał korę motoryczną pacjentów, w wyniku czego wykonywali oni ruchy ręką. Pacjenci ci zaprzeczali, że to oni samodzielnie zdecydowali o ruchu ręki i wskazywali na Penfielda jako tego, który uruchomił określone działanie. Należałoby przyjąć, zgodnie z podejściem Searle'a, że gdyby pacjent Penfielda sam zamierzał podnieść rękę, ale jej ruch zostałby wywołany przez zewnętrzne pobudzenie elektryczne jego kory motorycznej, to osoba ta nie uznałaby ruchu swojej ręki za efekt własnego zamiaru, gdyż ruchowi temu nie towarzyszyłoby poczucie działania (*experience of acting*), którego treść pokrywa się z treścią intencji w działaniu. Innymi słowy, w sytuacji, w której pojawia się zamiar, ale wykonywany ruch nie jest odczuwany jako zamierzony, zamiar uznany zostaje za niezrealizowany. Wskazywałem, omawiając eksperyment Libeta oraz koncepcję Wegnera (por. rozdz. 4. niniejszej rozprawy), że problem sprawczości zamiaru jest znacznie bardziej złożony, niż zaproponowane przez Searle'a rozwiązanie.[88] Zawieszę tu tę kwestię, gdyż jej rozstrzygnięcie nie wpływa na postać modeli działania intencjonalnego, które proponuję niżej.

Rozważmy teraz drugą cechę intencjonalności zamiaru, jaką jest skierowanie ku cielesnym ruchom (zachowaniom), traktowanym jako szczególnego rodzaju przedmioty intencjonalne. Uznanie przez Searle'a, że to ruch ciała jest przedmiotem intencjonalnym zamiaru jest niekonwencjonalne, gdyż zgodnie ze standardowym pojmowaniem intencjonalności, przedmiotem takim jest zewnętrzny względem podmiotu (ujmowanego jako ciało wraz z jego stanami umysłowymi) obiekt lub stan otoczenia. Nawet gdyby pominąć nasuwające się filozoficzne wątpliwości, czy w świetle deklarowanego naturalizmu biologicznego ujęcie takie daje się utrzymać[89] i przyjąć, że w treści zamiaru rzeczywiście zawarte jest powiązanie z ruchem ciała, to pojawia się pytanie, czy Searle'owska koncepcja działania intencjonalnego wiążąca wąsko pojmowany zamiar[90] z ruchem ciała nie jest nadmiernie uproszczona. Być może uproszczenie takie pomaga w analizach czysto filozoficznych, ale w swojej oryginalnej postaci koncepcja ta jest na tyle enigmatyczna, że ze względu na jej ogólnikowość można dopasować ją do dowolnych,

[88] Mowa w tym miejscu głównie o tym, że eksperymentalnie wykazano, iż to nie zamiar wywołuje ruch, lecz występujące jeszcze przed pojawieniem się zamiaru procesy (tzw. potencjał gotowości) w systemie motorycznym mózgu (Libet 1983).

[89] Jeśli w zgodzie z naturalizmem biologicznym przyjąć, że zamiar jest stanem (wyższym) mózgu, to pojawia się pytanie: po co mózgowi intencjonalne, a więc w pewnym sensie zdalne, kierowanie ręką, skoro może on bezpośrednio wpływać na jej ruch poprzez łączące go z nią fizyczne „okablowanie"?

[90] Rozumiany jako intencja w działaniu.

zarówno prostych, jak i złożonych działań intencjonalnych. Wystarczy włączyć do tła (por. omówienie Searle'owskiego pojęcia tła w rozdziale 2. niniejszej pracy) niemieszczące się w wąsko pojmowanym zamiarze cechy standardowo rozumianego działania intencjonalnego, by można było za ich pomocą charakteryzować dowolnie wybrane ludzkie zachowania. Kiedy jednak podejmuje się próbę modelowania faktycznych ludzkich działań, to okazuje się, że mają one znacznie bardziej złożoną strukturę, niż postulował to Searle i trudno lokować ich niemieszczące się w zamiarze cechy w nieokreślonym bliżej tle. Kwestię tę rozważam poniżej, kiedy dyskutować będę odróżnienie między prostym a złożonym działaniem intencjonalnym. Tu skupię się na wykazaniu, że twierdzenie, iż przedmiotem intencjonalnym zamiaru jest pojmowany czysto biologicznie i charakteryzowany ogólnikowo ruch ciała, nadmiernie upraszcza strukturę działania intencjonalnego.

Zilustrujmy to na podstawie przywoływanego tu wielokrotnie przykładu podnoszenia ręki. Zgodnie z przedstawionym przez Searle'a opisem[91] działanie takie składa się z intencji w działaniu, jaką jest doznanie pojawiające się w trakcie podnoszenia ręki oraz z ruchu ciała podmiotu, który ma to doznanie. Brak doznania w sytuacji, kiedy ręka się podnosi, interpretowany jest jako ruch kończyny spowodowany przez czynniki niezależne od stanów umysłowych podmiotu. Natomiast wystąpienie doznania podnoszenia ręki (*experience of acting*), któremu nie towarzyszy jej ruch ku górze, traktowane jest jako przejaw nieskuteczności zamiaru. Skupmy się teraz na unoszeniu się ręki do góry i rozważmy, co rozumiał Searle przez określenie „ruch mojej ręki" (*the movement of my arm*)? Zgodnie z jego propozycją – ruch ten jest przedmiotem intencjonalnym, do którego

[91] „For the sake of simplicity, I will start with very simple actions such as raising one's arm. […] When I raise my arm I have a certain experience, and like my visual experience of the table, this arm-raising experience has a certain form of Intentionality, it has conditions of satisfaction. For if I have this experience and my arm doesn't go up, the Intentional content of the experience is not satisfied. Furthermore, even if my arm goes up, but goes up without this experience, I didn't raise my arm, it just went up. That is, just as the case of seeing the table involves two related components, an Intentional component (the visual experience) and the Intentional «object» or conditions of satisfaction of that component (the table), so the act of raising my arm involves two components, an Intentional component (the experience of acting) and the Intentional «object» or conditions of satisfaction of that component (the movement of my arm)." (Searle, 1980, s. 52, 55).
„Dla uproszczenia rozpocznę od bardzo prostych działań, takich jak podniesienie ręki. […] Kiedy podnoszę rękę, mam określone doznanie i – podobnie jak moje wzrokowe doznanie stołu – to doznanie podnoszącej się ręki ma określoną formę intencjonalności, ma warunki spełniania. Ponieważ, jeśli mam to doznanie, a moja ręka nie podnosi się w górę, intencjonalna treść doznania nie jest spełniona. Co więcej, nawet jeśli moja ręka podnosi się w górę, lecz dzieje się to bez doznania, to nie ja podniosłem rękę, to ona właśnie się uniosła. Znaczy to, że podobnie jak przypadek widzenia stołu zawiera dwa powiązane składniki: składnik intencjonalny (doznanie wzrokowe) oraz «przedmiot» intencjonalny, czyli warunki spełniające ten składnik (stół), tak czynność podnoszenia ręki zawiera dwa składniki: składnik intencjonalny (doznanie działania) oraz «przedmiot» intencjonalny, czyli warunki spełniające ten składnik (ruch mojej ręki)."

odnosi się składnik intencjonalny, jakim jest doznanie działania. Zastanówmy się, o jakim ruchu cielesnym tu mowa. Czy o dającym się stwierdzić obiektywnie (np. przez zarejestrowanie zmiany położenia ręki za pomocą kamery), a więc niezależnie od podmiotu intencji, unoszeniu się ręki ku górze? Czy raczej – skoro ma to być ruch *mojej* ręki – ruchu spostrzeganym przez podmiot intencji i z jego perspektywy?[92] Zwrot „ruch mojej ręki" sugeruje, że Searle nie miał na względzie przypadku pierwszego, a więc ruchu biologicznego w takiej postaci, w jakiej dostępny on jest zewnętrznemu obserwatorowi spostrzegającemu wzrokowo, że osoba, na której skupił swoją uwagę, unosi rękę. Nie ulega wątpliwości, że obserwator zewnętrzny patrzy na ruch cudzej ręki z zupełnie innej perspektywy, niż jej animator.[93] Także i drugie ujęcie – ruch ręki ujmowany z perspektywy podmiotu działającego – nie wydaje się w pełni zadowalające. Po pierwsze, kiedy podnoszę rękę, mam z jednej strony doznanie wzrokowe, z drugiej – doznanie kinestetyczne. Pomińmy tu obszerną problematykę wzrokowego spostrzegania ruchu własnej ręki[94] i spytajmy, czy kinestetyczne doznanie unoszącej się ręki nie jest tożsame z tym, co sam Searle określa jako doznanie działania? Taka interpretacja wydaje się nieuprawniona, gdyż doznanie kinestetyczne jest zespołem cielesnych odczuć pojawiających się jako rezultat procesu unoszenia ręki. Powiedzieć można, że podnosząca się ręka jest przyczyną doznań kinestetycznych, czyli to ona dostarcza bodźców, które odbierane są jako te właśnie doznania. Natomiast intencja w działaniu skierowana jest nie na minioną ani nawet nie na aktualną fazę ruchu, ale na tę fazę, która dopiero nastąpi. W przypadku doznania

[92] Można byłoby oponować przeciwko takiemu rozróżnieniu dwóch rodzajów ruchu ręki argumentując, że w istocie jest to jeden ruch, tyle – że ujmowany z dwóch różnych perspektyw: obiektywnej i subiektywnej. Problem polega na tym, że w przypadku ruchu ręki perspektywy te są niewspółmierne. Aby to uwidocznić, rozważmy znane z badań nad percepcją odróżnienie między egocentryczną a allocentryczną reprezentacją przestrzenną. W tej pierwszej relacje przestrzenne są zawsze relatywizowane do lokalizacji perceptora (traktowanej jako wyróżniony punkt odniesienia) oraz własności jego ciała. Natomiast w przestrzeni allocentrycznej nie ma wyróżnionego punktu odniesienia, a położenia spostrzeganych przedmiotów traktowane są jako niezależne od położenia perceptora (Klatzky, 1998). W przypadku ruchu ręki mówić można o perspektywie egocentrycznej, natomiast nie sposób patrzeć na ruch własnej ręki allocentrycznie, traktując go jako niezależny od reszty ciała (być może tak patrzą na ruchy własnej ręki osoby cierpiące na zespół obcej ręki). Mamy tu zatem do czynienia albo z perspektywą egocentryczną, albo perspektywą zewnętrznego obserwatora.

[93] Wyobraźmy sobie osobę, która odniosła poważną kontuzję barku. Leczenie wymagało wszczepienia implantu stawu barkowego i długotrwałej rehabilitacji, polegającej m.in. na ćwiczeniu unoszenia ręki do góry. To rehabilitant ocenia, czy ćwiczenia odniosły pożądany skutek i osoba rehabilitowana podnosi rękę dostatecznie wysoko. Rehabilitant może zachęcać podopieczną do wyższego podniesienia ręki. Ta jednak może oświadczyć, że osiągnęła pułap swoich możliwości i uznać, że to, co w oczach rehabilitanta jest niepełnym podniesieniem ręki, dla niej jest zupełnie wystarczające.

[94] Mowa tu przede wszystkim o tym, czy w trakcie podnoszenia ręki system wzrokowy działa w trybie percepcyjnym, czy w trybie wzrokowej kontroli działania. (Milner, Goodale 2008). Zgodnie z ustaleniami Milnera i Goodale'a należałoby przyjąć, że podczas podnoszenia ręki w górę przede wszystkim zaangażowany jest system kontroli wzrokowej, który działa poza świadomą kontrolą agenta.

traktowanego jako intencja w działaniu - to ono jest przyczyną ruchu: intencja w działaniu decyduje o zainicjowaniu całego procesu zmian położenia ręki, jak i o pojawianiu się kolejnych faz ruchu, a także o tym, kiedy ruch ten zostanie zakończony.

Zwróćmy uwagę na jeszcze jedną trudność. Otóż Searle pisze, charakteryzując intencję w działaniu, że kierunek dopasowania (*direction of fit*) ma w przypadku tego stanu umysłowego postać świat-do-umysłu.[95] Znaczy to, że intencja w działaniu, która – jak przeważająca większość stanów intencjonalnych – reprezentuje swoje warunki spełniania, zostanie wykonana (*carry out*), jeśli stan świata będzie dopasowany do stanu umysłu. W tym przypadku stanem umysłu (*mind*) jest zamiar danego podmiotu, natomiast stanem świata (*world*), który miałby być do niego dopasowany, jest odpowiadający treści zamiaru przedmiot intencjonalny: podnosząca się do góry ręka tego podmiotu. Ryzykowne wydaje się założenie mówiące o tym, że podmiot podejmujący działanie intencjonalne, polegające na wprawieniu w ruch części własnego ciała, będzie traktować zmiany w ich położeniu jako zmiany we fragmentach świata. Oznaczałoby to, że istota taka jest albo bezcielesna (to sam zamiar, niemający związku z ciałem, wprawia je w ruch), albo „wydziela" ona z własnego ciała „część", którą traktuje jako zewnętrzny (należący do świata) obiekt, na który następnie oddziałuje. Choć rozwikływanie tych i podobnych im problemów może być interesujące z perspektywy filozoficznej, skupionej na problemie umysł – ciało (*mind – body problem*), to jednak rozważania takie w nikłym stopniu pomagają w odsłonięciu struktury i mechanizmu działania intencjonalnego. Zastanówmy się zatem, co należałoby uczynić, aby uniknąć wskazanych wyżej trudności. Uważam, że pomocne okaże się poszerzenie zaproponowanej przez Searle'a dwuskładnikowej struktury działania intencjonalnego.

Działanie intencjonalne jako skierowanie ku stanowi mającemu wartość dla podmiotu działania, który to stan nie pojawi się bez zrealizowania tego działania

Zawieśmy chwilowo rozważania o Searle'a koncepcji działania intencjonalnego i zastanówmy się, jakie składniki zawierać powinna podstawowa charakterystyka takiego działania.

[95] „[…] in the case of the experience of acting, the Intentional component has the world-to-mind direction of fit." (Searle, 1980, s. 55).

Zacznijmy od tego, że działanie intencjonalne ma charakter celowy. Znaczy to, że agent zamierza osiągnąć za pośrednictwem działania stan (zwykle jest to stan świata wywołany przez cielesne zachowanie, ale niekiedy jest to także sam stan ciała agenta), który jest dla niego bardziej wartościowy, niż ten, w którym się aktualnie znajduje. Agent powinien więc potrafić wartościująco oszacować zastany stan otoczenia oraz zaplanować taką zmianę swojego zachowania, aby doprowadzić do wystąpienia stanu bardziej dla niego wartościowego. Ocena pod względem wartości stanu zastanego wymaga zebrania o nim danych, a osiągnięcie stanu zamierzonego wymaga posiadania planu działania. Trudno sobie wyobrazić, aby nie mając informacji o swoim położeniu oraz nie mogąc ocenić stanu zastanego agent zdecydował się na podjęcie działania. Co więcej, jeśli agent rozpoznaje stan, w jakim się znajduje i ocenia go jako w pełni satysfakcjonujący, to również nie podejmie działania. I wreszcie, jeśli agent rozpoznaje swoje położenie i ocenia je jako niesatysfakcjonujące, to podejmie działanie tylko wtedy, kiedy uzna, że potrafi je zrealizować, co w efekcie poprawi jego położenie. Podsumujmy: 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.

Odnieśmy powyższą charakterystykę działania intencjonalnego do rozważanego wyżej przykładu podnoszenia ręki. Otóż, aby podnoszenie ręki można było uznać za działanie intencjonalne, należałoby przyjąć, że agent rozpoznaje, iż ma opuszczoną rękę oraz stan, w którym jest ona podniesiona jest dla niego bardziej wartościowy, niż ten, w którym jest ona opuszczona. Agent chce osiągnąć stan bardziej dla niego wartościowy, jakim jest podniesiona ręka.

Jak pamiętamy, w ujęciu Searle'a sytuacja ta sprowadza się do tego, że agent chce (ma intencję w działaniu) podnieść rękę, co na mocy przyczynowości mentalnej prowadzi do tego, że unosi się ona w górę. Uważam, że taka charakterystyka jest zbyt uboga. Aby działanie agenta mogło zostać uznane za intencjonalne, potrzebna jest charakterystyka bogatsza, uwzględniająca jego wykonalność oraz spodziewaną wartość dla agenta. Kiedy agent chce zerwać wiszące wysoko nad jego głową jabłko, to nie zacznie podskakiwać z

wyciągniętą w górę ręką, jeśli oszacuje, że wisi ono zbyt wysoko by mógł je chwycić. O kimś, kto w takiej sytuacji zacząłby podskakiwać wiedząc, że nie ma szans na dosięgnięcie jabłka, nie powiemy, że chce on zerwać jabłko. Nie uznamy takiego podskakiwania za działanie intencjonalne, którego celem jest zerwanie jabłka.

Podsumujmy: skuteczne wykonanie działania intencjonalnego wymaga dysponowania przez agenta mechanizmem identyfikowania oraz waloryzowania stanów otoczenia (środowiska), a także mechanizmem wyznaczania celu, czyli wyszukiwania lub projektowania stanu bardziej wartościowego, niż zastany. Mechanizmy te opisane zostaną dokładniej w modelu 1. i modelu 2. działania intencjonalnego.

Działanie zrutynizowane a działanie z wbudowanym procesem uczenia się

Rozważmy teraz przypadek, w którym agent chce zerwać jabłko i szacuje, że podskakując z wyciągniętą w górę ręką powinien je dosięgnąć. Początkowo skacze pionowo w górę, stojąc pod zwisającym jabłkiem. Po kilku nieskutecznych próbach uznaje, że zdoła zerwać jabłko, jeśli skoczy po nie z rozbiegu. Dopiero taka zmiana zachowania przynosi zamierzony efekt. Zauważmy, że w tej sytuacji osiągnięcie celu wymagało szeregu prób i zmiany zachowania (nowego ruchu ciała). Powiemy, że agent osiągnął cel, gdyż po kilku niepowodzeniach nauczył się zachowania, które zapewnia mu osiągnięcie zaplanowanego, cennego dla niego stanu. Rozważany tu przypadek, w którym uwzględniony został mechanizm uczenia się, a więc wielokrotne powtarzanie i modyfikowanie czynności po to, aby osiągnąć zamierzony cel, bardzo rzadko analizowany jest w pracach poświęconych działaniom intencjonalnym. Zwykle, kiedy mowa o takim działaniu, bierze się pod uwagę działanie z jednorazowym zachowaniem, a więc takie, w którym cel zostaje osiągnięty już po pierwszym wykonaniu ruchu cielesnego. Łatwo zauważyć, że w takim przypadku abstrahuje się od funkcji, jaką pełni mechanizm uczenia się w działaniu intencjonalnym. W istocie, trudno wyobrazić sobie działanie intencjonalne, które podejmowane byłoby w pełni spontanicznie, bez uczenia się. W realnych sytuacjach, w których mamy do czynienia z działaniami wykonywanymi w trybie jednorazowego zachowania, cielesne ruchy zawdzięczają swoją skuteczność temu, że zostały wyuczone wcześniej. Kiedy w trakcie obiadu sięgamy widelcem po leżący na talerzu kawałek brokułu i płynnie wkładamy go do ust, wykonujemy sekwencję czynności, których nauczyliśmy się w dzieciństwie. Gdybyśmy, zamiast widelca, trzymali w ręku chińskie pałeczki, z którymi nie mieliśmy

wcześniej do czynienia, to prosta czynność przeniesienia kawałka brokułu z talerza do ust z pewnością nie przebiegłaby tak płynnie. Działaniami zrutynizowanymi nazwijmy działania intencjonalne wykonywane wyłącznie za pomocą wcześniej wyuczonych zachowań. Wyuczone wcześniej mogą być zarówno pojedyncze zachowania, jak i całe ich sekwencje. W tym ostatnim przypadku uczenie się dotyczy zarówno jednostkowych zachowań, jak i porządku ich wykonywania. Kwestia, w jaki sposób przebiega mechanizm wytwarzania działań zrutynizowanych, warta jest odrębnej analizy. Z mojej perspektywy działanie zrutynizowane jest działaniem prostym. Nabrało ono skrystalizowanej formy na skutek wcześniejszego, skutecznego wyuczenia się, ale w jego aktualnym wykonaniu mechanizm uczenia się nie jest uruchamiany. O działaniu takim nie powiemy, że jest inteligentne, choć z pewnością posiada ono wskazane wyżej cechy działania intencjonalnego. Przedmiotem mojego zainteresowania są działania intencjonalne z wbudowanym, wewnętrznym mechanizmem uczenia się. Ze względu na to, iż mechanizm uczenia się jest ich nieusuwalnym składnikiem, proponuję nazwać je złożonymi działaniami intencjonalnymi. Zwykle ich składnikami są działania proste, czyli zrutynizowane, ale są one wkomponowane w strukturę, w której działa mechanizm wyboru optymalnych zachowań w trybie uczenia się. Działanie, które stopniowo podwyższa swoją skuteczność, jawić się będzie zewnętrznemu obserwatorowi jako inteligentne.

Wielopoziomowa struktura działania intencjonalnego

Największym wyzwaniem poznawczym – przy tak sformułowanym celu badawczym – jest problem zintegrowania ze sobą niezbędnych do realizacji działania danych oraz konstruktów teoretycznych pochodzących z różnych poziomów opisu działań intencjonalnych, w szczególności trudność ta polega na pokazaniu związku między wysokopoziomowym opisem funkcjonowania sieci stanów intencjonalnych z niskopoziomowymi mechanizmami uczenia się ze wzmacnianiem. Każdy z wymienionych poziomów opisu korzysta z zasadniczo odmiennego aparatu pojęciowego. Searle'a teoria intencjonalności, przy pomocy której scharakteryzowana została sieć stanów intencjonalnych, odwołuje się do pojęcia stanu intencjonalnego S(p) złożonego z określonej treści oraz modusu psychologicznego. Z kolei opis mechanizmu uczenia się ze wzmacnianiem opiera się na mechanistyczno-obliczeniowej ramie pojęciowej. Trudno zaklasyfikować istniejącą między nimi relację jako prostą hierarchię, w której jeden poziom w pełni wpływa na drugi. Należy również podkreślić, że wzajemne oddziaływanie

stanów intencjonalnych oraz mechanizmu uczenia się ze wzmacnianiem zmienia się w czasie. Ocenia się, że kontrola zachowań na podstawie wzmocnień jest szczególnie cenna dla osób młodych – do dwudziestego roku życia (Shephard i in., 2014). W tym okresie stopniowo wykształcana jest umiejętność planowania, która z czasem zaczyna odgrywać coraz większą rolę.

> *Istniejące badania zdają się potwierdzać intuicyjne oczekiwanie, że stosunkowo proste zadania planowania opanowywane są przez dzieci już w klasach początkowych. Natomiast, umiejętność realizacji bardziej złożonych zadań tego typu pojawia się dopiero w klasach średnich oraz w okresie dojrzewania.*[96]

To strategiczne przesunięcie w radzeniu sobie z rzeczywistością, jak się wydaje, silnie wiąże się z uwagą Dretskiego o szczególnej wadze wiedzy teoretycznej (przekonań łączących) w postrzeganiu rzeczywistości, a w konsekwencji – również w działaniu. Dysponując tego typu wiedzą uzyskuje się wyższą trafność predykcji, ale również wydłuża się jej horyzont czasowy. Wskazana dynamika – stopniowe zwiększanie udziału planowania w procesie doboru zachowań – stanowi istotne wymaganie względem zintegrowanego modelu działań intencjonalnych. W dalszej części rozdziału lista tego typu wymagań zostanie uzupełniona o – wskazane przez Patricka Haggarda – cechy działań intencjonalnych.

Sformułuję, tworząc listę uznanych za niezbędne cech działań intencjonalnych, składniki eksplanandum, dla którego szukanym eksplanansem jest sekwencja proponowanych modeli. Do konstrukcji takiego eksplanansu proponuję zastosować schemat stosowany powszechnie w informatyce, gdzie proces wytwarzania oprogramowania dzieli się – na wysokim poziomie abstrakcji – na następujące fazy:

1. faza definiowania i analizy wymagań,
2. faza konstrukcji i implementacji systemu,
3. faza testowania systemu[97].

[96] „The existing research, then, seems to validate the intuitive expectation that relatively simple planning tasks are mastered by school-aged children in the early grades. Performance on more complex planning tasks, however, continues to develop beyond middle childhood and through adolescence" (Parrila i in., 1996, s. 598).

[97] Przyjąłem, że faza testowania systemu – ze względu na teoretyczny charakter prowadzonych w pracy rozważań – zostanie pominięta w niniejszej dysertacji.

Powyższy schemat wpływa na strukturę kolejnych części rozdziału. Na początku – w formie zestawu najistotniejszych cech – uzupełniona i uszczegółowiona zostanie lista najważniejszych wymagań wobec zintegrowanego modelu działań intencjonalnych. Następnie zaprezentowane zostaną poszczególne komponenty modelu działań intencjonalnych.

5.2 Cechy złożonego działania intencjonalnego

Na podstawie analiz przeprowadzonych w rozdziałach 2., 3. i 4. można wskazać kilka podstawowych cech systemu, którym jest złożone działanie intencjonalne. Aby wpisać tego rodzaju projekt w konteksty wcześniej przeprowadzonych analiz, najpierw przypomniane zostaną najważniejsze rezultaty badań przedstawionych w poprzednich rozdziałach.

Podejście filozoficzne charakteryzuje się z pewnością najszerszym zakresem analizy, odnosząc się do przypadków prostych (np. pociągnąć za spust), podstawowych (np. wykonać strzał) oraz złożonych (np. pomścić Serbię)[98]. Szczególnie istotne w tym podejściu jest wpisanie działań intencjonalnych w szerszą ramę teoretyczną, jaką jest Searle'owska teoria intencjonalności. Jej krytyczną analizę przedstawiłem wyżej. Tu wskażę na zawarte w niej idee, które wykorzystane zostaną przy konstruowaniu modelu. Teoria ta wiąże działania z dwoma istotnymi konstruktami: (1) siecią stanów intencjonalnych, stanowiącą zaawansowaną bazę wiedzy podmiotu oraz (2) przedintencjonalnymi dyspozycjami tła, będącymi rezerwuarem zautomatyzowanych umiejętności oraz przedintencjonalnych sposobów odnoszenia się do środowiska. W tak zdefiniowanym kontekście możliwe staje się zidentyfikowanie typów stanów intencjonalnych oraz ich wewnętrznej struktury. Prior intencja i intencja w działaniu to, zdaniem Searle'a, stany umysłowe w szczególny sposób związane z zachowaniami. Pozwalają one odróżnić spontaniczne i planowane działania dobrowolne.

Obydwa typy intencji nigdy nie funkcjonują w izolacji. Ich treść zawsze powiązana jest z szerszym kontekstem, np. z określonymi pragnieniami, przekonaniami, lękami, obawami, nadziejami, itd. (patrz: idea holizmu znaczeniowego[99]). Tego rodzaju sieć nadbudowana jest nad przedintencjonalnymi umiejętnościami, nastawieniami i dyspozycjami tła, które wyznaczają horyzont możliwości dla konstruowanych stanów intencjonalnych.

[98] Porównaj rozdział 2. Przyczynowy wpływ stanów intencjonalnych na wybór zachowań.
[99] Patrz rozdział 2. Przyczynowy wpływ stanów intencjonalnych na wybór zachowań, str. 38.

Zachowania „kontrolowane" przez stany intencjonalne mogą mieć postać prostych, jednostkowych ruchów lub złożonych sekwencji, które z czasem, w wyniku powtórzeń, mogą uzyskać status umiejętności tła, czyli czynności wysoce zautomatyzowanych. Tego typu umiejętności, które nabywane są na ogół metodą prób i błędów, można, zdaniem Searle'a, traktować jako szczególnego rodzaju reprezentacje. Pozbawione są one treści, jednak – oprócz złożonych programów motorycznych – zapewniają one bogaty informacyjnie układ odniesień do otaczającego nas świata (np. do trwałości obiektów fizycznych, ich ciężaru, faktury, itd.), umożliwiając w ten sposób pojawienie się stanów intencjonalnych.

Inną kwestią, na którą zwraca uwagę amerykański filozof, jest czasowy wymiar działań intencjonalnych. Sekwencja zdarzeń, w opinii Searle'a, musi mieć ściśle określony przebieg w tego typu działaniach. W przeciwnym przypadku działanie nie zostanie uznane za zamierzone. Każde istotne zaburzenie schematu postulowanego przez Searle'a prowadzi do zmiany statusu działania na nieumyślne, nieplanowane czy wręcz niechciane. Spostrzeżenie Searle'a potwierdza efekt scalania (*binding effect*), zbadany eksperymentalnie m.in. przez Patricka Haggarda (Haggard, 2005).

Wymienione elementy teorii intencjonalności uwzględnione są w zintegrowanym modelu działań intencjonalnych jako dwa współpracujące ze sobą podsystemy: (1) podsystem zarządzania siecią stanów intencjonalnych oraz (2) podsystem nadzorowania i realizacji celów. Pierwszy odpowiedzialny jest za reprezentowanie oraz konstrukcję planów, w tym prior intencję, rola drugiego polega na aktywowaniu wskazanego przez prior intencję celu oraz na doborze i kontroli działań niezbędnych do jego osiągnięcia. Kolejnym składnikiem modelu jest podsystem uczenia się ze wzmacnianiem (więcej na ten temat w dalszej części rozdziału).

Wykorzystywane przez neurobiologów podejście obliczeniowe modeluje złożone zachowania celowe za pomocą algorytmów uczenia się ze wzmacnianiem (np. algorytm TDRL). Algorytmy te wymagają do swojego działania dobrze zdefiniowanych zachowań elementarnych (np. dobrowolnych ruchów ciała) oraz co najmniej dwóch typów danych, z których każdy odpowiada stosownemu rodzajowi reprezentacji umysłowej. Dane te reprezentują: (1) stany środowiska zarejestrowane za pomocą procesów percepcyjnych oraz (2) nagrody wyrażone w formie informacji o tym, jakie korzyści uzyska agent „odwiedzając" poszczególne stany środowiska. Tego typu mechanizm – za pomocą

odpowiednich reguł obliczeniowych – stopniowo przekształca zachowania realizowane metodą prób i błędów w działania celowe, pozyskujące nagrody w niemal optymalny sposób. Wskazana cecha tego typu algorytmów umożliwia agentowi pozyskanie dostępnych w środowisku zasobów przy minimalnym nakładzie kosztów. Innymi słowy, ujęcie komputacyjne pozwala wyjaśnić złożone działania celowe za pomocą odpowiednio zorganizowanej polityki doboru zachowań elementarnych (tzw. akcji) oraz określonych, wysokopoziomowych cech algorytmu, takich jak efekt blokowania[100] czy zdolność do przedkładania nagród długoterminowych nad krótkoterminowymi. Przyczynowość w algorytmie TDRL ma charakter w pełni mechanistyczny, tzn. poszczególne kroki realizowane są w ściśle określony sposób i to niezależnie od tego czy środowisko, w którym funkcjonuje agent, jest deterministyczne i stacjonarne, czy cechuje się nieusuwalną zmiennością. Choć algorytmy należące do rodziny RL nie wymagają do swojego działania złożonych stanów intencjonalnych, to – jak pokazał Read Montague – pojęcie nagrody w nich stosowane jest na tyle pojemne, że bez większych problemów może ono objąć złożone umysłowe stany intencjonalne takie jak pragnienia, przekonania, idee, itp., umożliwiając wyjaśnienie za pomocą mechanizmu uczenia się ze wzmacnianiem tak problematycznych od strony biologicznej przypadków, jak głodówki z przyczyn politycznych czy zbiorowe samobójstwa w sektach religijnych. Inną cechą algorytmu uczenia się ze wzmacnianiem jest również otwartość na rozszerzenia. Wśród dostępnych rozszerzeń warto wspomnieć o (a) możliwości hierarchizacji zachowań (idea tzw. opcji stanowiących uogólnienie zachowań elementarnych) oraz (b) zdolności do wykorzystywania wiedzy proceduralnej, dostarczanej agentowi w formie nagród kształtujących (za pomocą tego typu nagród można np. przekazać agentowi informację o obecności ściśle określonej liczby nagród w danej przestrzeni, ograniczając w ten sposób zakres eksplorowanej przestrzeni).

Przedstawiona powyżej charakterystyka mechanizmu uczenia się ze wzmacnianiem (w pełni zaprezentowana w rozdziale 3.) wykorzystana została do skonstruowania jednego z najważniejszych podsystemów zintegrowanego modelu działań intencjonalnych, mianowicie podsystemu hierarchicznego uczenia się ze wzmacnianiem z optymalizacją domenową. Jest on odpowiedzialny za realizację dwóch ważnych funkcji: (1) za integrowanie zachowań elementarnych w złożone sekwencje stanowiące bazę dla działań

[100] Efekt blokowania polega na tworzeniu asocjacji pomiędzy bodźcem poprzedzającym nagrodę a nagrodą. W określonych przypadkach tego typu sekwencja asocjacji może obejmować bardzo wiele bodźców (Montague, 2006, s. 108)

intencjonalnych oraz (2) za dostarczanie podsystemowi zarządzania siecią stanów intencjonalnych informacji niezbędnych do jego rozszerzenia. Psychologowie intencji (m.in. Patrick Haggard, Daniel Wegner, Benjamin Libet) w swoich badaniach również odwołują się do prior intencji oraz do intencji w działaniu (choć nie zawsze używają dokładnie tych wyrażeń). Z ich perspektywy, wymienione stany to nie tylko konstrukty teoretyczne, ale przede wszystkim obiekty, które można badać metodami empirycznymi. Obecnie neuropsychologom i neurofizjologom udało się w dużym stopniu rozpoznać wewnętrzną strukturę intencji w działaniu. Stan ten, wbrew temu, co twierdzi Searle, nie oddziałuje na zachowania w sposób przyczynowy, ale im towarzyszy – posiada status korelatu świadomościowego dla procesów motorycznych. Można w nim wyróżnić dwie składowe: subiektywno-motoryczną (chęć wykonania ruchu) oraz sensoryczno-celowościową (odniesienie do docelowego obiektu lub zdarzenia) (Gomes, 1998; Haggard, 2005). Ponadto, psychologowie intencji wnioskują na podstawie swoich badań o występowaniu nieświadomych procesów odpowiedzialnych za dobór oraz realizację zachowań celowych (patrz: eksperyment Haynesa oraz czasowy przebieg potencjału gotowości w eksperymencie Libeta) (Soon i in., 2008). Udało się również rozpoznać efekt scalenia czasowego składowych działania intencjonalnego, który nie jest obecny w przypadku działań mimowolnych. Innym, ważnym wnioskiem dotyczącym tego obszaru badań jest rozpoznanie własności poczucia sprawstwa, czyli fenomenu odpowiedzialnego za reprezentowanie zachowań jako skutków naszych świadomych decyzji. Pojawienie się tego szczególnego rodzaju doznania jest rezultatem złożonego procesu, który – w zależności od kontekstu – ma charakter post-rekonstruktywistyczny (patrz: iluzyjna koncepcja świadomej woli Daniela Wegnera, autora *The illusion of conscious will*) lub predykcyjny (patrz: eksperymenty Haggarda oraz ujęcie Fritha (Frith, 2012)).

Przytoczone powyżej wyniki badań, szczegółowo zaprezentowane w rozdziale 4., stanowią bazę dla jednej z głównych hipotez leżących u podstaw zintegrowanego modelu działań intencjonalnych. Zgodnie z tą hipotezą procesy odpowiedzialne za wystąpienie poczucia sprawstwa są przejawem procesu, który przekształca niskopoziomowe reprezentacje wykorzystywane przez podsystem uczenia się ze wzmacnianiem w wysokopoziomowe reprezentacje złożone, czyli stany intencjonalne (więcej na ten temat znajduje się w sekcji omawiającej najbardziej złożoną wersję zintegrowanego modelu działań intencjonalnych, tj. model 3.0).

Zestawienie wyników badań przywołanych powyżej pokazuje, że nie są one jednorodne. Różnice dotyczą m.in.:

- poziomu opisu (ogólny, wysokopoziomowy opis Searle'a a opracowana przez psychologów intencji drobiazgowa analiza składowych intencji),
- wykorzystywanego aparatu pojęciowego (teoria intencjonalności a model obliczeniowy oparty na algorytmie TDRL) oraz
- zakresu badań (analiza złożonych zachowań Montague'a a identyfikacja składowych prostych działań intencjonalnych w badaniach psychologicznych).

Różnice dotyczą również sposobów interpretowania zjawisk (koncepcja przyczynowości intencjonalnej Searle'a a iluzyjna hipoteza świadomej woli Daniela Wegnera). Uważam, że warto – mimo wskazanych różnic między poszczególnymi koncepcjami – kierować się zasadą konsiliencji (Wilson, 2002), a tam, gdzie pojawiają się rozbieżności, promować rozwiązanie najbardziej prawdopodobne z perspektywy całościowego modelu.

Etapy konstrukcji sekwencji modeli, z których każdy reprezentuje układ złożony z wymienionych wyżej systemów, prezentowane są zgodnie ze stosowaną w informatyce metodyką wskazaną na początku niniejszego punktu. W związku z tym omówienie poszczególnych modeli poprzedzone zostanie krótką analizą cech, które powinny zostać przy ich pomocy wyjaśnione. Podstawą do ich wyznaczenia będą ustalenia trzech badaczy: Patricka Haggarda, Johna Searle'a oraz Reada Montague'a.

Cecha 1: Zależność od kontekstu i wyuczonych wcześniej asocjacji

> *Działania intencjonalne tylko w niewielkim stopniu zależą od bezpośrednich bodźców, zaś w dużym stopniu zależą od kontekstu zadania oraz od wyuczonych wcześniej powiązań.* (Haggard, 2005, s. 291).

Ta cecha wskazuje, iż u podstaw działań intencjonalnych znajduje się złożony mechanizm kontroli zachowań, który wykracza poza prosty schemat „bodziec→reakcja behawioralna" (tak, jak to ma miejsce w przypadku odruchów). Odpowiednio ukierunkowany mechanizm uczenia się, który stopniowo optymalizuje wybór zachowań, umożliwia reakcje agenta na zmieniające się okoliczności tak, aby były one uzgodnione z wcześniejszymi

doświadczeniami. Działania pozbawione tego typu ukierunkowania byłyby przypadkowe, w gruncie rzeczy sprowadzałyby się do spontanicznych reakcji na aktualnie odbierane bodźce, a więc nie podlegałyby wyuczonym prawidłowościom zachowań ani nabytej wcześniej wiedzy o własnościach otoczenia. Nie trzeba dodawać, że takie spontaniczne działania byłyby nieefektywne, a ze względu na brak poprzedzającego je celu (prior intencja) albo chociaż intencji w działaniu trudno byłoby określić je jako intencjonalne (patrz: faza manii w chorobie dwubiegunowej, która cechuje się natłokiem skojarzeń i myśli). Warto zauważyć, że mechanizm uczenia się ze wzmacnianiem, który zorientowany jest na stopniową optymalizację zachowań, odsłania niewidoczną na pierwszy rzut oka złożoność działań. Zakłada się błędnie, że zapamiętanie informacji o efektach zachowania wystarcza, by decydent niejako automatycznie, w przyszłości, poprawił efektywność działań nakierowanych na podobny cel. Tymczasem jest to jedynie warunek wstępny, co pokazują prace z obszaru uczenia maszynowego oraz robotyki (Mnih i in., 2015; Sutton, 1998). Podmiot działający – bez eksploracji środowiska, bez stosowania predykcji, bez ciągłego monitorowania długoterminowych rezultatów oraz odpowiednich sposobów aktualizacji dotychczas stosowanej strategii doboru zachowań – nigdy nie uzyska znaczącej poprawy jakości swoich działań, nawet gdyby systematycznie powiększał się zasób jego asocjacji. Choć Patrick Haggard[101] nie określa typu asocjacji, które wpływają na przebieg działań intencjonalnych, to warto w tym przypadku przyjąć najszerszą z możliwych interpretacji, czyli założyć, że asocjacje odnoszą się zarówno do sądów intencjonalnych (np. przekonań, pragnień, wcześniejszych zamiarów), jak i do percepcji oraz przedintencjonalnych nastawień (tzw. umiejętności tła). O ile udział asocjacji w sądach i działaniach intencjonalnych nie budzi większych wątpliwości (utwierdza nas w tym refleksja odnosząca się do informacji o wpływie tego typu stanów na nasze wybory i działania), o tyle trudniej jest uwzględnić w nich np. umiejętności tła. W tym przypadku pomocne są badania dotyczące uczenia warunkowego zwierząt oraz psychologii rozwojowej dzieci (Olds, 1958; Shephard i in., 2014) . Niezależnie od tego – czy funkcję determinant pełnią stany intencjonalne (sądy lub percepcje), czy przedintencjonalne dyspozycje tła – łatwo zauważyć, że odbierane bodźce podlegają redeskrypcji ze względu na umysłowe reprezentacje miejsc czy sytuacji, co w konsekwencji wpływa na wybór sekwencji zachowań. Towarzyszące zachowaniom – na danym etapie rozwoju dziecka –

[101] Działania intencjonalne tylko w niewielkim stopniu zależą od bezpośrednich bodźców, zaś w dużym stopniu zależą od kontekstu zadania oraz od zapamiętanych wcześniej asocjacji (Haggard, 2005, s. 291).

stany intencjonalne (głównie stany percepcyjne i przypomnienia) mogą nie być dostępne w formie jawnych przekonań (patrz: trudności z zaliczeniem testu fałszywych przekonań u dwu- i trzyletnich dzieci (Reuter, 2014)), a mimo to mogą determinować przebieg zachowań. Dwa wymiary funkcjonowania asocjacji wskazane przez Haggarda (zależność od kontekstu oraz od wyuczonych wcześniej powiązań) można odnieść – w nawiązaniu do koncepcji Searle'a – do dwóch głównych składowych teorii intencjonalności, mianowicie do tła (składowa niskopoziomowa obsługiwana m.in. przez mechanizm uczenia się ze wzmacnianiem) oraz do elementów sieci stanów intencjonalnych (asocjacje związane z prior intencją, intencją w działaniu, pragnieniami, przekonaniami, itd.). W kontekście tego rozróżnienia jeszcze wyraźniej widać, że sposób reprezentowania informacji napływających do systemu kontroli zachowań jest wielowymiarowy i wymaga odpowiednich procesów poznawczych oraz leżących u ich podstaw procesów obliczeniowych. Reprezentowanie informacji w przypadku procesów niskopoziomowych odbywa się głównie poza naszą świadomością, natomiast posługiwanie się intencjami oraz innymi stanami intencjonalnymi wymaga dodatkowych mechanizmów oraz zasobów poznawczych, dlatego ważne jest uwzględnienie kolejnej cechy działań intencjonalnych, którą wskazał Patrick Haggard.

Cecha 2: Udział sieci procesów poznawczych w planowaniu i kontroli działania intencjonalnego

> *Przygotowanie i wykonanie działań intencjonalnych może wymagać skupienia uwagi, a rezultaty działań bywają monitorowane przez procesy poznawcze w związku z uczeniem się na przyszłość.* (Haggard, 2005, s. 291).

Ta cecha wskazuje na silny związek działań intencjonalnych z procesami poznawczymi oraz tzw. funkcjami wykonawczymi (uwaga, pamięć robocza, percepcja, planowanie, monitorowanie, poznawcza elastyczność (Carlson i in., 2005; por. Jodzio, 2008, s. 44)), które są niezbędne do konstruowania, a następnie wykorzystywania sieci stanów intencjonalnych do kontroli zachowań. Funkcjonowanie wymienionych procesów pozwala na dalszą optymalizację działań intencjonalnych. Model świata zawarty w sieci stanów intencjonalnych, który stopniowo jest poszerzany w trakcie rozwoju ontogenetycznego, otwiera przed agentem zupełnie nowe możliwości. Możemy, rozumiejąc jak działa określone zjawisko w świecie, zorganizować tak nasze zachowania, aby uniknąć

kosztownych, a często również niebezpiecznych błędów, np. rozpoznać sygnały zbliżającego się tsunami, przechodzić na drugą stronę ulicy tylko w wyznaczonych do tego miejscach, itp. Czasochłonna i zasobochłonna systematyczna eksploracja środowiska stosowana w metodzie uczenia się ze wzmacnianiem zostaje w ten sposób radykalnie ograniczona. Koszt, w formie deliberacji i planowania, który przychodzi nam w związku z tym ponieść, jest istotnie mniejszy w stosunku do czasochłonnej i często ryzykownej eksploracji. Oczywiście, zgromadzona przez nas wiedza nie zawsze musi być adekwatną reprezentacją rzeczywistości. Czasami zawarte w przekonaniach informacje na temat funkcjonowania danego zjawiska w świecie mogą prowadzić do pogorszenia efektów działania, a w skrajnych przypadkach (np. fałszywe przekonania, błędne plany, głupota[102]) mogą całkowicie pozbawić agenta szansy na osiągnięcie celu. Pomimo tego mankamentu, nadal przewaga organizmów dysponujących tego typu modelem świata oraz procesami poznawczymi, konstruującymi na jego podstawie predykcje dotyczące przyszłych stanów świata, jest znacząca w porównaniu z organizmami pozbawionymi takich możliwości[103].

Najwyraźniej wpływ procesów poznawczych i funkcji wykonawczych na kształtowanie działań można dostrzec, porównując zachowania ludzkie do zachowań zwierzęcych. Badania psychologii porównawczej (Trojan, 2013) wyraźnie pokazują, że zwierzęta, w szczególności naczelne, dysponują podstawowymi formami planowania oraz zdolnościami rozwiązywania problemów, jednak „pojemność" (*capacity*) tych dyspozycji w zasadniczy sposób odbiega od tego, czym dysponuje człowiek. Język, rozumiany jako złożony system komunikacyjny,[104] jest niewątpliwie jednym z kluczowych czynników wpływających na zwiększanie naszych zdolności planowania i deliberacji, które decydują

[102] Zdaniem Davida Krakauera, teoretyka systemów złożonych (*complex systems*), głupota rozumiana jako strategia poszukiwania rozwiązania jest znacząco gorsza od strategii losowej – to jedno z największych wyzwań, przed którym stoi współczesny świat (Paulson, 2015).

[103] Koncepcja istot popperowskich Daniela Dennetta opiera się na podobnym spostrzeżeniu. W ocenie amerykańskiego filozofa zdolność do symulacji, zdolność przewidywania przyszłych stanów świata, stawianie hipotez, rozstrzyganie, które z zaplanowanych działań będą korzystne dla organizmu, a które nie – stanowi ważne osiągnięcie ewolucyjne (D. C. Dennett, 1997).

[104] Zgodnie z ujęciem Charlesa Hocketta, za język można uznać taki system komunikacyjny, który posiada następujące cechy:

- posługiwanie się odpowiednim kanałem informacyjnym,
- arbitralność,
- semantyczność,
- przekaz kulturowy,
- spontaniczność,
- dialogowość,
- dwoistość strukturalna,
- strukturalność,
- autonomiczność mowy,
- kreacyjność (za: Trojan, 2013, s. 32).

o złożoności naszych zachowań oraz o horyzoncie czasowym, który mogą one objąć. To z jego pomocą procesy tworzenia nowych reprezentacji oraz manipulowania nimi zyskują jakościowo odmienny charakter w porównaniu z analogicznymi procesami u zwierząt[105].

Ludzka zdolność do zaawansowanego przetwarzania informacji zapewnia człowiekowi przewagę nad innymi gatunkami zwierząt, jednak koszty związane z podtrzymywaniem i aktywizowaniem tego typu zdolności nie są, jak twierdzi Montague, niskie, dlatego ich wykorzystywanie musi podlegać rozmaitym ograniczeniom. W określonych sytuacjach może się bowiem okazać, że z perspektywy organizmu mniej kosztowne, a bardziej użyteczne od strony realizacji celu, będzie zastosowanie metody prób i błędów, niż deliberacji i planowania opartych na złożonych procesach poznawczych. Współwystępowanie wymienionych metod kontroli zachowań – uczenia się ze wzmacnianiem oraz planowania - wymaga wypracowania sposobów ich współpracy i koordynacji. Jest to jedno z ważniejszych zagadnień, które podjęte zostanie podczas konstrukcji zintegrowanego modelu złożonych działań intencjonalnych. Z zagadnieniem tym wiąże się również kolejna cecha wyróżniona przez Haggarda.

Cecha 3: Udział deliberacji i planowania w wyborze sposobu realizacji działania

> *Wybór działań intencjonalnych na ogół jest poprzedzany przez procesy planowania i deliberacji, które wymagają wysiłku poznawczego*. (Haggard, 2005, s. 291).

Powyższa cecha działań intencjonalnych odnosi się głównie do przypadków, w których ważną funkcję pełni plan opracowany jako rezultat procesu deliberacji. Prior intencja jest zapowiedzią takiego planu. Pomiędzy prior intencją a działaniami intencjonalnymi *sensu stricto* zachodzi, zgodnie z analizą Searle'a[106], związek „treściowo" – przyczynowy, który determinuje na odpowiednio wysokim poziomie wybór zachowań oraz relacje między nimi. Można zatem przyjąć, że prior intencja pełni w systemie kontroli zachowań rolę „kotwicy", za pomocą której podmiot ma dostęp do właśnie opracowanego planu działania

[105] „Na dziesięć wymienionych w definicji Charlesa Hocketta cech, siedem pierwszych nie jest specyficznych wyłącznie dla ludzkiego sposobu komunikowania się. Za unikalne przynależne naszemu językowi do dziś można wciąż uznać jedynie trzy ostatnie: strukturalność, autonomiczność i kreacyjność. W takim ujęciu używanie terminu «język» w przypadku komunikowania się innych gatunków należałoby potraktować jako nadużycie." (Trojan, 2013, s. 39).

[106] Patrz: „Schemat przebiegu działania intencjonalnego wg Searle'a" opisany w podrozdziale 2.2. *Schemat pełnego działania intencjonalnego*.

(por. Jodzio, 2008, s. 38). Utworzenie prior intencji wymaga na ogół odpowiedniej analizy, rozważenia dostępnych opcji realizacji celu oraz nałożenia na to własnych preferencji i doświadczeń. Tego typu proces wykorzystuje na ogół odpowiednie zasoby, w skład których wchodzą: funkcje wykonawcze oraz zaawansowane procesy poznawcze (m.in. abstrahowanie, podejmowanie decyzji, rozwiązywanie problemów, formułowanie sądów czy planowanie (por. Nęcka i in., 2006a)). Działania podejmowane na podstawie wymienionych funkcji i procesów są konfrontowane, podobnie jak niskopoziomowe mechanizmy uczenia się ze wzmacnianiem, z uzyskiwanymi rezultatami. Bez oceny rezultatów nie mogłyby się rozwijać mechanizmy: samokontroli oraz poznawczej elastyczności, a co za tym idzie – nie byłaby możliwa realizacja coraz bardziej złożonych działań. Z perspektywy mechanizmu monitorującego oznacza to, że zaobserwowane rezultaty działań powinny zostać odwzorowane w dwóch niezależnych repozytoriach: (1) repozytorium umiejętności, z którego „korzysta" mechanizm uczenia się ze wzmacnianiem oraz (2) repozytorium stanów intencjonalnych (patrz: sieć stanów intencjonalnych Searle'a). Tego typu „architektura" umożliwia wzajemną kontrolę obu sposobów reprezentowania rzeczywistości, a co za tym idzie – ich modyfikację w sytuacji, gdy uzyskiwane rezultaty są niespójne. Oczywiście, taki układ wymaga sprawnie działających mechanizmów synchronizacji i to zarówno w krótkim, jak i w długim horyzoncie czasowym. Innym zagadnieniem, które wiąże się z trzecią cechą działań intencjonalnych, jest wbudowany w nie mechanizm uczenia się oraz konstruowania nowych reprezentacji świata. Obserwacje rozwoju psychomotorycznego dzieci wyraźnie pokazują, że takie umiejętności jak: siedzenie, zmiana pozycji z jednej na drugą, zdolność stania, chodzenia, itd., okupione są dziesiątkami nieskutecznych prób. Podobnie, jak się wydaje, przebiega proces poznawania najbliższego otoczenia oraz realizowania podstawowych potrzeb, takich jak zaspokojenie głodu czy pragnienia. Barton Sutton, jeden z ojców wielkiego sukcesu metody uczenia się ze wzmacnianiem, zauważa:

> *Idea, że uczymy się poprzez interakcję z naszym otoczeniem, jest prawdopodobnie pierwszą, która przychodzi nam do głowy, kiedy myślimy o naturze uczenia się. Kiedy niemowlę bawi się, macha rękami lub rozgląda się, nie ma wyraźnego nauczyciela, ale za to ma bezpośrednie połączenie sensoryczno-motoryczne z otoczeniem.*

> *Korzystanie z tego połączenia daje mnóstwo informacji o przyczynach i skutkach, o konsekwencjach działań i o tym, co należy zrobić, aby osiągać cele*[107].

Z czasem umiejętności nabywane metodą prób i błędów stają się coraz bardziej złożone i zaczynają funkcjonować jako podstawowe jednostki w procesach planowania i deliberacji.

O ile mechanizm uczenia się ze wzmacnianiem jest w pełni gotowy do działania już od chwili narodzin człowieka, o tyle procesy deliberacji i planowania wymagają wieloletniego treningu i to w środowisku, w którym będą w przyszłości wykorzystywane. Wyraźnie pokazują tę zależność od wyuczonych norm i zachowań choćby systemy prawne, w których rozróżnia się czyny popełnione przez dzieci, nastolatków oraz przez osoby dorosłe. Dzieci poniżej 13. roku życia – zgodnie z obowiązującymi w wielu państwach demokratycznych kodeksami karnymi – nie podlegają karze, a za ich ewentualne przewinienia odpowiadają opiekunowie. Z kolei, w przypadku czynów zabronionych popełnianych przez nieletnich (od 13. do 17. roku życia), stosuje się, poza wyjątkami, inne przepisy, niż w przypadku analogicznych czynów popełnianych przez osoby dorosłe (Jurgielewicz-Delegacz, 2019). Zdolność do przewidywania konsekwencji własnych działań jest jedną z kluczowych umiejętności, którą musi wykształcić młody człowiek, zanim będzie mógł w pełni odpowiadać za swoje czyny. Ta forma predykcji – w odróżnieniu od mechanizmu uczenia się ze wzmacnianiem – nie służy do pozyskiwania nagród. Jej głównym zadaniem jest ocena moralnego i społecznego aspektu działania ze względu na skutki, które może wywołać. Mechanizm kontroli zachowań oparty jedynie na karach i nagrodach jest zbyt prosty, dlatego też wymaga odpowiednich rozszerzeń, by uwzględniona została złożona sieć relacji, jaka funkcjonuje w obrębie przekonań odnoszących się do kulturowo zdeterminowanych norm społecznych.

Warto zauważyć, iż wykorzystanie obu metod kontroli zachowań tj. (1) planowania działań oraz (2) uczenia się ze wzmacnianiem, zmienia się w czasie. Najpierw działania intencjonalne organizowane są przez mechanizmy uczenia się ze wzmacnianiem, z czasem jednak kontrola stopniowo przejmowana jest przez mechanizmy planowania i deliberacji.

[107] „The idea that we learn by interacting with our environment is probably the first to occur to us when we think about the nature of learning. When an infant plays, waves its arms, or looks about, it has no explicit teacher, but it does have a direct sensorimotor connection to its environment. Exercising this connection produces a wealth of information about cause and effect, about the consequences of actions, and about what to do in order to achieve goals" (Sutton, 1998, s. 4).

Tego typu reorientacja nie jest oczywiście całkowita, w związku z tym do listy cech działań intencjonalnych należy dodać kolejną.

Cecha 4: Zastępowanie prostszych mechanizmów uczenia się przez mechanizmy bardziej zaawansowane

Powyższa cecha wskazuje na dynamikę obecną w systemie odpowiedzialnym za realizację działań intencjonalnych, która polega na zmianie sposobu funkcjonowania mechanizmu doboru zachowań. Początkowo mechanizm ten bazuje na metodzie uczenia się ze wzmacnianiem, by następnie przybrać postać podejścia opartego w głównej mierze na „prowizorycznym" planowaniu. Wyrażenie: „prowizoryczne planowanie" służy zasygnalizowaniu problematycznego statusu pojęcia „plan". Zmienność i nieprzewidywalność środowiska przyrodniczego, jak twierdzi Read Montague, w istotny sposób ogranicza przydatność detalicznego planu działania. Znacznie bardziej efektywne jest rozwiązanie, w którym planowanie traktuje się jedynie jako element usprawniający metodę uczenia się ze wzmacnianiem. Warto w tym miejscu zauważyć, że obserwowana w świecie przyrody zmienność i nieprzewidywalność ograniczana jest cywilizacyjnie. Różnego rodzaju procedury, regulacje, zasady postępowania, reguły stosowane w procesach opisujących funkcjonowanie organizacji (m.in. finansowych, politycznych, przemysłowych) stabilizują całe środowisko naturalne (lub przynajmniej redukują jego dynamikę) i znacząco przyczyniają się do zmniejszenia jego nieprzewidywalności (Searle, 1995a). W dużym stopniu zanika więc, opisana przez Montague'a, wada szczegółowego, niekorygowalnego planowania.

Kontrola zachowań oparta na dwóch mechanizmach: (1) planowaniu oraz (2) uczeniu się ze wzmacnianiem nakłada jeszcze jedno wymaganie na cały układ tworzący działanie intencjonalne. Mianowicie wymaga koordynacji oraz synchronizacji poszczególnych etapów procesu umożliwiającego realizację pełnego działania intencjonalnego. W tym miejscu należy przypomnieć, że związki treściowe, które istnieją pomiędzy stanem intencjonalnym a zachowaniem, nie wystarczają do tego, aby dane działanie dobrowolne można było uznać za zamierzone. Poszczególne stany umysłu oraz ruchy muszą być odpowiednio skoordynowane w czasie, aby podmiot uznał dane zachowanie za zgodne z jego wcześniejszą intencją (por. omówione wcześniej eksperymenty Penfielda). Jest to niewątpliwie jeden z ważniejszych warunków stawianych modelowi działania intencjonalnego.

Przedstawione powyżej cztery najważniejsze cechy działań intencjonalnych potraktowane zostaną w kolejnej sekcji niniejszego rozdziału jako kluczowe wymagania dla zintegrowanego modelu. Przyjęto, ze względu na złożoność tego modelu, że jego poszczególne składowe będą wprowadzane stopniowo – wraz z opisem czterech cech działań intencjonalnych.

5.3 Zintegrowany model złożonego działania intencjonalnego

Zintegrowany model złożonego działania intencjonalnego zaprezentowany zostanie przy wykorzystaniu układu podsystemów, które są potrzebne do realizacji omówionych wcześniej wymagań. Poszczególne podsystemy służą realizacji ściśle określonych funkcji i zarządzaniu specyficznymi dla nich reprezentacjami. Aby cały układ mógł efektywnie działać, podsystemy muszą ze sobą współpracować, realizując wynikające z ich funkcji zadania w kontekście procesów składających się na system realizacji działań intencjonalnych. Współpraca pomiędzy poszczególnymi komponentami modelu zaprezentowana zostanie w formie relacji istniejących między poszczególnymi podsystemami.

Podczas prezentacji modelu wykorzystywane będą pojęcia rozumiane w następujący sposób:

- System – skoordynowany układ funkcjonalnych podsystemów, realizujący proste i złożone działania intencjonalne. Struktura systemu wyznaczona przez podsystemy jest zasadniczo niezmienna, jednak efekty działania podsystemów zmieniają się w czasie. Dzieje się tak z powodu zmian będących skutkami ich rozwoju.
- Podsystem – zbiór wyspecjalizowanych modułów realizujący wysokopoziomowe funkcje, które współorganizują przebieg działania intencjonalnego. Podsystem charakteryzuje się specjalizacją dziedzinową, która wynika z architektury całego systemu (por. Miłkowski, 2009, s. 40). Zakłada się przy tym, że nadzorowane przez podsystem reprezentacje są przez niego w pełni kontrolowane, zarówno ich stan, jak i dostępność dla innych podsystemów (porównaj: idea izolacji informacyjnej Fodora (*encapsulation*) (Miłkowski, 2009, s. 30)).
- Moduł – elementarna jednostka obliczeniowa służąca realizacji ściśle określonej, niskopoziomowej funkcji w ramach danego podsystemu; przyjęte podejście wzorowane jest na zaproponowanej przez Sternberga koncepcji modułu, dla

którego „moduł to po prostu niezależna jednostka o osobnej funkcji; jej podstawowym wyróżnikiem jest możliwość zmiany w izolacji od reszty podsystemu (*separate modifiability*)" (por. Miłkowski, 2009, s. 31)); przyjęta definicja – w porównaniu z podejściem zaproponowanym przez Jerry'ego Fodora – „nie wymaga" od modułu specjalizacji dziedzinowej (*domain specificity*), izolacji informacyjnej (*informationally encapsulated*), szybkości przetwarzania ani nawet określonej neuronalnej lokalizacji (por. Miłkowski, 2009, s. 31).

- Reprezentacja – struktura **analogiczna** do struktury danych w informatyce, która jest przedmiotem przekształceń w obrębie procedur obliczeniowych (algorytmów) stanowiących **analogon** procesów mentalnych (por. Thagard, 2020). W pracy przyjmuję, że reprezentacje funkcjonujące w poszczególnych podsystemach mogą posiadać formę zarówno struktur symbolicznych, jak i struktur rozproszonych, wykorzystywanych do modelowania sztucznych sieci neuronowych.

Inne pojęcia techniczne, które są wykorzystywane w modelu, będą definiowane sukcesywnie – wraz z jego prezentacją.

Proponowane podejście będzie się odbywało etapami, zgodnie z metodą kolejnych przybliżeń. Etapy te zorganizowane zostaną wokół zdefiniowanych wcześniej cech działań intencjonalnych. Pełna architektura złożonego działania intencjonalnego wyłoni się stopniowo. Dany etap prezentacji modelu przebiegać będzie w trzech fazach. W fazie pierwszej pokazane zostanie, jak dana cecha działania intencjonalnego wykorzystana jest w konstrukcji modelu. Następnie zaprezentowany zostanie diagram obrazujący podsystemy, które są niezbędne do realizowania funkcji implikowanych przez wskazaną cechę. Wreszcie, w ostatniej fazie przedstawiona zostanie argumentacja potwierdzająca przydatność modelu w wyjaśnianiu postawionych przed nim wymagań.

5.3.1 Model 1.0 – wpływ kontekstu i wyuczonych asocjacji na strukturę i przebieg złożonego działania intencjonalnego

Przyjmuję za Haggardem (por. sekcja 5.2, Cecha 1.), że istotną cechą działania intencjonalnego jest zależność od kontekstu i wyuczonych wcześniej powiązań. Znaczy to, że podstawowa charakterystyka takiego działania uwzględniać powinna to, że nie jest ono bezpośrednią reakcją na bodźce odebrane z otoczenia, lecz inicjowane jest samodzielnie przez podmiot, który w procesie uczenia się nabył szereg umiejętności oraz wiedzę niezbędną do jego wykonania. Charakterystyka ta powinna również uwzględniać to, że

podmiot aktywnie eksploruje otoczenie i szacuje (na podstawie wyuczonych umiejętności oraz wiedzy), że zamierzone działanie jest wykonalne w rozpoznanych przez niego warunkach, a stan będący rezultatem działania będzie bardziej wartościowy, niż stan aktualny.

Wyuczone powiązania dotyczyć mogą zarówno niskopoziomowych reprezentacji stanów świata, nagród oraz zachowań (patrz: metoda uczenia się ze wzmacnianiem), jak i reprezentacji wyższego poziomu (patrz: różnego rodzaju obiekty lub stany abstrakcyjne). Przyjmując, że rozbudowany system działań intencjonalnych wykorzystuje obydwa typy reprezentacji, należy uznać, że w swojej strukturze zawierać on będzie co najmniej dwa wyspecjalizowane podsystemy. Ich prezentacja, ze względu na ich złożoność, realizowana będzie stopniowo w kilku odsłonach (wersjach).

Pojęcie powiązania odniesione zostanie najpierw do reprezentacji niskopoziomowych, których „konsumentem" oraz „producentem" będzie podsystem uczenia się ze wzmacnianiem (Model 1.0). Następnie, podsystem ten zostanie połączony z innymi podsystemami, tak aby cały układ pozwalał wyjaśnić zachowania stosunkowo złożonych organizmów biologicznych (Model 1.1). W kolejnym kroku, wybrane podsystemy zostaną wyposażone w możliwość generalizacji nadzorowanych przez nie reprezentacji (Model 1.2). Proces generalizacji obejmie obserwacje (o/O), nagrody (r/R) oraz zachowania (z/Z). Ostatni z typów wymaga dodania do modelu nowego podsystemu, którego główną funkcją jest konstruowanie zachowań złożonych będących sekwencjami zachowań prostych. Takie podejście pozwala w sposób przyrostowy zaprezentować najważniejsze składowe modelu spełniającego wymagania właściwe dla Cechy 1. Należy podkreślić, że inspiracją dla takiej konstrukcji systemu kontroli zachowań, spełniającego pierwsze wymaganie Haggarda, jest hipoteza dopaminergicznego błędu predykcji nagrody oraz jej model obliczeniowy. Odpowiednio rozszerzona wersja algorytmu TDRL (*Temporal Difference Reinforcement Learning)*, zgodnie z przedstawionym poniżej uzasadnieniem, pozwala wyjaśnić na poziomie behawioralnym te działania, które opierają się na wyuczonych powiązaniach.

Model 1.0 - charakterystyka działania intencjonalnego ograniczona do powiązań niskopoziomowych

Najbardziej wyidealizowany, uwzględniający powiązania niskopoziomowe, układ podsystemów, decydujący o podjęciu działania intencjonalnego, można zobrazować za pomocą następującego diagramu:

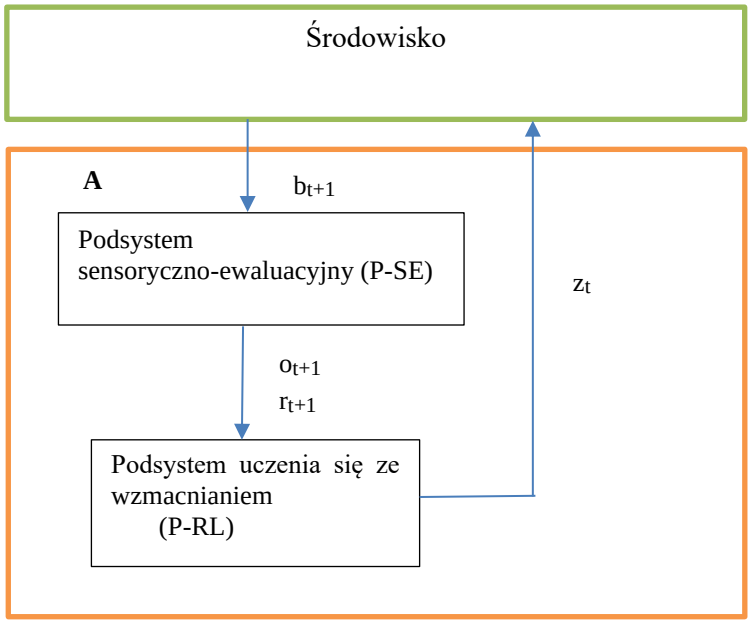


Diagram 10. Model działania intencjonalnego uwzględniający wyłącznie powiązania niskopoziomowe wersja 1.0.

Legenda użytych symboli:

- z_t – zachowanie zrealizowane przez agenta w chwili t; za wybór zachowania z_t odpowiada podsystem P-RL;
- b_{t+1} – bodźce odebrane przez agenta w chwili t+1, wygenerowane przez środowisko, będące w stanie s_t w związku z zachowaniem z_t, np. b_{t+1} to sygnały świetlne docierające do siatkówki oka odbierane w związku z wykonaniem określonego ruchu głowy;
- o_{t+1} – obserwacja o_{t+1} jest reprezentacją utworzoną przez podsystem sensoryczno-ewaluacyjny (P-SE); odnosi się do stanu środowiska w chwili ‘t+1’; podstawą do utworzenia tego typu reprezentacji są bodźce b_{t+1}, które są reakcją środowiska na zachowanie z_t, którego realizacji podjął się agent, np. agent tworzy odpowiednią reprezentację percepcyjną na podstawie pobudzenia układu wzrokowego, a następnie rozpoznaje – na podstawie określnych cech tej reprezentacji – że znalazł się w stanie s_t;
- r_{t+1} – nagroda r_{t+1} to pochodząca ze środowiska natychmiastowa zwrotna informacja wartościująca, oceniająca stan ‘s_t’ z perspektywy realizowanego celu; za utworzenie tego typu reprezentacji odpowiedzialny jest określony moduł w podsystemie sensorycznym (P-SE), który – na podstawie docierających do agenta

pobudzeń sensorycznych – wyznacza ich bieżącą wartość, np. wartość pożywienia określona jest na bazie sygnałów pochodzących z układu węchowo-smakowego.

Uzasadnienie

Diagram 10 prezentuje interakcje między agentem a środowiskiem. Środowisko wpływa na agenta poprzez określone bodźce, natomiast agent oddziałuje na środowisko za pomocą dostępnych mu zachowań. Głównym wyzwaniem dla agenta jest dobór zachowań, który będzie optymalny ze względu na dany stan środowiska oraz cel, który pragnie osiągnąć, wyrażony jako chęć pozyskania określonego typu nagród. Agent, aby mógł wykonać tego typu optymalizację, powinien dysponować podsystemem sensoryczno-ewaluacyjnym (P-SE) oraz podsystemem uczenia się ze wzmacnianiem (P-RL). Pierwszy podsystem, czyli P-SE w następujących po sobie jednostkach czasu (t), na podstawie napływających bodźców (b_{t+1}) oraz na podstawie zgromadzonych wcześniej doświadczeń – dostarcza podsystemowi P-RL reprezentacje dwojakiego rodzaju: (1) obserwacje (o_{t+1}) oraz (2) nagrody (r_{t+1}) (są one związane z bieżącym stanem środowiska). Pierwszy typ reprezentacji pozwala podsystemowi P-RL określić stan środowiska, w którym agent właśnie się znalazł (s_{t+1}). Drugi typ reprezentacji odnosi się do nagrody, czyli natychmiastowej zwrotnej informacji wartościującej. Tego typu informacja pozwala agentowi ocenić, na ile stan środowiska jest dla niego korzystny lub niekorzystny ze względu na obrany cel. Wymienione reprezentacje służą agentowi do wyboru następnego zachowania 'z_{t+1}' adekwatnego do postawionego przed agentem celu. Opisaną sekwencję kroków – z perspektywy wykorzystywanych przez podsystem reprezentacji – przedstawić można w następujący sposób:

$o_1, r_1 \rightarrow z_2 \rightarrow o_2, r_2 \rightarrow z_3 \rightarrow \ldots \rightarrow o_n, r_n \rightarrow z_{n+1}$.

Podsystem P-RL opiera swoje działanie na algorytmie uczenia się ze wzmacnianiem z zastosowaniem metody różnic czasowych[108] (TDRL – *temporal difference reinforcement learning*). Algorytm ten zapewnia, że prowadzone przez agenta obserwacje oraz realizowane zachowania prowadzą stopniowo do opracowania skutecznej (niemal

[108] „Metody różnic czasowych są pewnego rodzaju niestandardowym podejściem do wieloetapowych problemów predykcyjnych. W takich problemach należy na każdym etapie wygenerować prognozę pewnej nieznanej końcowej wartości na podstawie dostępnej w tym kroku cząstkowej informacji. Można przyjąć, że w kolejnych krokach informacja ta jest coraz bardziej pełna i wiarygodna, powinna więc umożliwiać lepsze stawianie prognozy. W trakcie uczenia się predykcje generowane w poszczególnych krokach modyfikuje się za pomocą błędów obliczanych jako różnice wartości przewidywanych w dwóch kolejnych krokach czasu, w jednym, którego dotyczy modyfikacja, oraz następnym, w którym prognoza przez domniemanie powinna być lepsza." (P. Cichosz, 2007, s. 754).

optymalnej) strategii doboru działań z uwzględnieniem zasady, że „dobra strategia nie od razu musi przynieść dobre efekty, ale [powinna] sprawdzić się w dłuższym horyzoncie czasowym" (P. Cichosz, 2007, s. 717). TDRL stanowi dla agenta gwarancję, że za każdym razem, gdy nastąpi trwała zmiana środowiska, wpływająca na efektywność dotychczasowego zachowania (pojawią się nowe przeszkody lub dotychczas pozyskiwane nagrody zmienią swoje położenie lub nawet znikną), to dojdzie do odpowiedniej korekty strategii doboru zachowań. Warto w tym miejscu przypomnieć, że w algorytmie TDRL wykorzystuje się metodę prób i błędów, a to znaczy, że nie wymaga się od agenta znajomości środowiska (*free-model* RL), gdyż jego struktura i dynamika „odkryte" zostaną w trakcie realizowanych interakcji.

Inną ważną cechą podsystemu P-RL, o której należy wspomnieć w analizowanym kontekście, jest jego „otwartość" na możliwość uczenia się zarówno w trybie: „z załączoną polityką doboru zachowań" (*on-policy*), jak i w trybie: „bez załączonej polityki" (*off-policy*). Pierwszy tryb prowadzi do stopniowej poprawy efektywności doboru zachowań, a z czasem do wypracowania strategii optymalnej, poprzez sukcesywne korygowanie funkcji wartości V, reprezentującej dla stanu świata 's' możliwą do pozyskania z tego stanu zdyskontowaną sumę przyszłych nagród. Drugi tryb – odpowiadający uczeniu się przez naśladownictwo – pozwala opanować zupełnie nowe formy zachowań na podstawie obserwacji innych agentów. W takim przypadku, zaobserwowane działania (akcje) traktowane są jak wzorce, które należy wiernie odtworzyć w danej sytuacji. Połączenie wymienionych trybów zdecydowanie przyspiesza proces nabywania złożonych umiejętności, w szczególności zastosowanie trybu bez załączonej polityki (*off-policy*) pozwala znacznie skrócić etap „tworzenia" działań złożonych z zachowań prostszych. W formach treningu realizowanych w sztukach walki (np. w karate) znaleźć można interesującą analogię. Z jednej strony adepci powtarzają wielokrotnie te same ruchy, będące coraz bardziej sformalizowanymi układami ćwiczeń, doskonaląc przy tym techniki (tzw. *kihon*) oraz ich sekwencje (*kata*) (tryb *off-policy*). Z drugiej strony nabyte umiejętności wykorzystuje się podczas sparingów (*kumite*), kiedy to próbuje się wykorzystać wyuczone wcześniej techniki do odparcia lub wyprowadzenia ataku podczas konfrontacji z przeciwnikiem (tryb *on-policy*).

Przedstawiona powyżej wersja modelu 1.0 działań intencjonalnych spełnia w wąskim zakresie charakterystykę Cechy 1. działań intencjonalnych. System, w którym

wykorzystywane są jedynie niskopoziomowe[109] asocjacje (powiązania), oparte na obserwacjach, nagrodach i zachowaniach, jest zbyt prosty, by można było go uznać za zadowalające wyjaśnienie złożonego działania intencjonalnego. Warto jednak podkreślić, że uwzględnia on te aspekty Cechy 1., bez których pełniejsza charakterystyka ludzkiego, złożonego działania intencjonalnego nie jest w ogóle możliwa. W takim modelu nie ma miejsca na prior intencję, intencję w działaniu, pragnienia, przekonania czy ich obliczeniowe bądź neuronalne korelaty. Znaczy to, że mechanizm realizowania celowych zachowań wyłącznie z pomocą algorytmu TDRL jest funkcjonalnie zbyt prosty, aby można było potraktować go jako podstawę podejmowania złożonych działań intencjonalnych. Twierdzę, że przedstawiony powyżej układ dwóch podsystemów jest adekwatnym wyjaśnieniem funkcjonowania nie tylko wybranych typów sztucznych systemów, korzystających z metod uczenia się ze wzmacnianiem (robotów, sztucznych operatorów gier komputerowych, itp.), ale także prostszych organizmów biologicznych, np. takich jak ślimak *Aplysia californica*. Powyższy model nie wystarcza do opisu zachowań wyższych zwierząt np. ssaków dysponujących zdolnością uczenia się warunkowego (tzw. pawłowowski tryb uczenia się). Brakuje w nim bowiem podsystemu, który aktywowałby lub dezaktywował aktualnie realizowany cel. Zwierzęta (kręgowce, a w szczególności ptaki czy ssaki) są agentami „wielocelowymi", które muszą spełniać różnego rodzaju potrzeby. Do najbardziej podstawowych potrzeb w ich przypadku zalicza się: zaspokojenie głodu i pragnienia, potrzebę reprodukcji i bezpieczeństwa. Znaczy to, że model w wersji 1.0 wymaga rozbudowania i włączenia w niego dodatkowych podsystemów.

5.3.2 Model 1.1 – działanie intencjonalne z podsystemem kontroli celów i podsystemem projektowania ich zmiany

W modelu 1.0 uwzględnione zostały podsystemy i zależności występujące na niższych poziomach funkcjonowania podmiotu działającego. W systemie opisywanym przez Model 1.0 pojęcia „wiedza" czy „doświadczenie" nie mogą być rozumiane literalnie. Warto jednak zwrócić uwagę na to, że skoro o przywołanym wyżej ślimaku *Aplysia californica*

[109] W pracy rozróżniam dwa rodzaje powiązań (asocjacji): (1) niskopoziomowe oraz (2) wysokopoziomowe. Pierwsze rozpatrywane są wyłącznie jako związki, do których agent nie ma i nie może mieć świadomego dostępu, a tym bardziej świadomej nad nimi kontroli. Powiązania wysokopoziomowe z kolei odnoszą się do stanów intencjonalnych oraz do relacji między nimi, czyli funkcjonują na poziomie tzw. sieci stanów intencjonalnych. W przypadku pierwszych powiązań mówimy zatem o mechanizmach molekularnych i neuronalnych, które nie wymagają odniesienia do wyższych stanów umysłowych, a tym bardziej świadomości.

mówi się, że jest zdolny do uczenia się i przechowywania informacji w formie pamięci długotrwałej, to jest to równoważne z przypisaniem mu zdolności do gromadzenia swoistej dla tego gatunku wiedzy i doświadczeń. Dlatego też uważam, iż zgodnie z Modelem 1.0 na działanie intencjonalne w jego najprostszej wersji składają się: (a) odbiór bodźców z otoczenia, (b) ich waloryzacja, (c) „odniesienie" wyselekcjonowanych po waloryzacji bodźców do zgromadzonej wcześniej wiedzy i doświadczenia utworzonego przez podsystem uczenia się ze wzmacnianiem.

Naszkicowana tu struktura działania intencjonalnego nie uwzględnia dwóch niezwykle ważnych składników: (1) mechanizmu (podsystemu) homeostazy, pozwalającego ocenić, czy agent znajduje się w stanie umożliwiającym mu podjęcie działania oraz (2) mechanizmu (podsystemu) monitorowania i motywacji pełniącego rolę dodatkowego selektora bodźców, które przesyłane są do podsystemu uczenia się. Model 1.1 przedstawia układ z Modelu 1.0 wzbogacony o wymienione wyżej podsystemy.

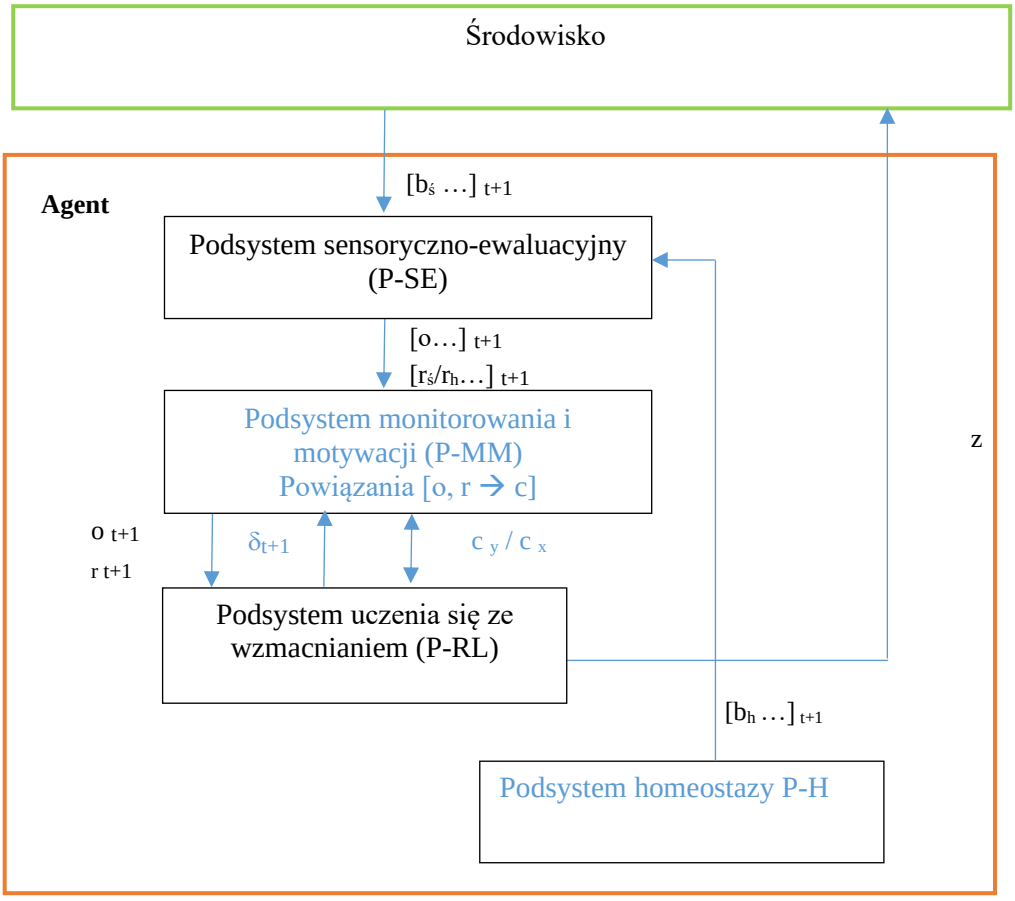


Diagram 11. Model działania intencjonalnego z podsystemem kontroli celów i podsystemem projektowania ich zmiany wersja 1.1.

Legenda symboli (uzupełnienie w stosunku do Modelu 1.0):

- $[b_{ś}\ldots]_{t+1}$, $[b_{h}\ldots]_{t+1}$ – bodźce pochodzące ze środowiska zewnętrznego lub wewnętrznego (podsystemu homeostazy), na podstawie których tworzone są reprezentacje obserwacji (o) oraz nagród ($r_{ś}$ i r_{h});
- $[o\ldots]_{t+1}$ – zbiór obserwacji ‘o’ odnoszących się do bieżącego stanu środowiska, utworzony na podstawie bodźców $b_{ś\ t+1}$ oraz $b_{h\ t+1}$; poszczególne obserwacje przekazywane są do podsystemu monitorowania i motywacji, którego głównym zadaniem jest rozpoznawanie obserwacji relewantnych z perspektywy realizowanego celu oraz ignorowanie obserwacji nieistotnych; dlatego po przejściu przez podsystem P-MM zbiór [o…] redukowany jest symbolicznie do pojedynczej obserwacji ‘o’, istotnej z perspektywy celu[110]; tego typu obserwacja, podobnie jak w wersji 1.0 modelu, umożliwia podsystemowi P-RL utworzenie wynikającej z niej reprezentacji stanu środowiska ‘s’;
- $[r_{ś}\ldots]_{t+1}$ – zbiór nagród reprezentujący natychmiastową zwrotną informację wartościującą na temat bieżącego stanu środowiska; wartość nagrody wyznaczana jest przez moduł ewaluacji zawarty w podsystemie P-SE, który „wycenia” napływające informacje, uwzględniając przy tym dane pochodzące z podsystemu homeostazy, tzn. odpowiednio zwiększa lub zmniejsza wartość nagrody $r_{ś}$ w zależności od tego czy organizm jest w stanie równowagi, czy jest zaburzony; podsystem P-MM „filtruje” dostępne nagrody, podobnie jak w przypadku zbioru [o…], udostępniając podsystemowi uczenia się ze wzmacnianiem wyłącznie nagrodę, która jest relewantna z perspektywy realizowanego celu;
- $[r_{h}\ldots]_{t+1}$ – nagroda ‘r_{h}’ jest utworzona na podstawie bodźców pochodzących z podsystemu homeostazy, reprezentuje natychmiastową informację zwrotną wartościującą, która odnosi się do bieżącego stanu organizmu; ten typ nagrody pozwala organizmowi realizować cele związane z zabezpieczeniem jego podstawowych potrzeb, w tym m.in. potrzebę bezpieczeństwa, bliskości, itp.; ponadto, ten typ nagród sygnalizuje podsystemowi monitorowania i motywacji przypadki naruszenia stanu homeostazy – np. braki energetyczne organizmu powodują pojawienie się stanu głodu odczuwanego jako nieprzyjemny;

[110] Wskazana w opisie reguła filtracji obserwacji ma charakter umowny. W praktyce, stosowany filtr może być mniej lub bardziej skuteczny, co oznacza, że w określonych przypadkach w podsystemie P-RL może być aktywowana więcej niż jedna tego typu reprezentacja.

- δ_{t+1} – błąd predykcji nagrody dla stanu s_{t+1} jest obliczany w ramach podsystemu uczenia się ze wzmacnianiem; służy do optymalizacji strategii doboru zachowań oraz informowania podsystemu monitorowania i motywacji o ewentualnych niedoszacowaniach lub przeszacowaniach danego stanu świata w odniesieniu do realizowanego celu ‘c’;
- c_x / c_y – operacja, która polega na dezaktywacji celu ‘x’ oraz na aktywacji celu ‘y’ w podsystemie P-RL; inicjatorem tego typu operacji jest podsystem monitorowania i motywacji, który na podstawie asocjacji typu „obserwacja-nagroda-cel” ([o,r,c]) decyduje o tym, kiedy – na podstawie informacji wartościującej ‘r’ lub obserwacji ‘o’ – należy aktywować cel ‘c’.

Uzasadnienie

Zaprezentowana na powyższym rysunku rozszerzona wersja modelu 1.0 odnosi się do wyróżnionej przez Patricka Haggarda pierwszej cechy działań intencjonalnych, tym razem jednak – na skutek zastosowania dodatkowych podsystemów (zaznaczonych na niebiesko) –wskazany model znacząco poszerza zakres swojego zastosowania. Przede wszystkim nowy model zmienia relację pomiędzy podsystemem sensomotoryczno-ewaluacyjnym a podsystemem uczenia się ze wzmacnianiem. W przedstawionym rozwiązaniu obserwacje oraz nagrody „wygenerowane” przez środowisko lub wewnętrzny stan agenta, zanim zostaną przekazane do podsystemu uczenia się ze wzmacnianiem, najpierw przetwarzane są przez podsystem monitorowania i motywacji (P-MM). P-MM wykorzystuje również, oprócz informacji pochodzących ze środowiska ([$b_{ś}$… $_{t+1}$]), sygnały pochodzące z podsystemu homeostazy (P-H) [$b_{h\ t+1}$]. Podsystem P-SE, korzystając m.in. z układu neuroendokrynnego i immunologicznego, dostarcza do P-MM informację o potrzebach organizmu wynikających z jego bieżącego stanu (np. kończących się zasobów energetycznych, przeżywanego stresu, itp.). W ten sposób środowiskowe informacje (np. dostrzeżony cień antylopy) uzyskują dodatkowy kontekst. Przykładowo, ten sam typ nagrody (np. pożywienie) będzie inaczej traktowane przez agenta, kiedy będzie on głodny, a inaczej – kiedy będzie syty; odmienna będzie reakcja na zagrożenie, gdy osobnik będzie w pełni sił, a inna – gdy będzie chory. Bardzo ważną składową całego podsystemu jest zatem element motywacyjny P-MM, który decyduje o priorytetach pozyskiwania określonego rodzaju nagród.

P-MM, uwzględniając rozszerzoną hipotezę bramkowania dopaminergicznego opracowaną przez Montague (patrz: podrozdział zatytułowany *Inicjowanie celów abstrakcyjnych* w rozdziale 2.), pełni dodatkowo dwie istotne funkcje. Pierwsza związana jest ze **stabilizacją** podsystemu P-RL polegającą na zapewnianiu podsystemowi uczenia się odpowiednio długiego czasu działania. W systemie wielocelowym występuje bowiem istotny z perspektywy przetrwania organizmu dylemat decyzyjny: czy należy kontynuować aktualnie realizowany cel, czy – po dostrzeżeniu nowego typu nagrody lub zaburzeniu homeostazy – przełączyć się na nowy cel, być może korzystniejszy? Sztucznie wywołana destabilizacja układu odpowiedzialnego za tego typu decyzje, jak pokazały eksperymenty Donalda Strussa i Roberta Knighta (Stuss & Knight, 2002), może prowadzić do fiksacji celu albo do zbyt częstego przełączania pomiędzy celami. Obydwie strategie prowadzą do poważnych zaburzeń, a w naturalnym środowisku mogą powodować nawet śmierć organizmu. Rozwiązanie wskazanego dylematu polega na czasowym odizolowaniu podsystemu P-RL od napływających nowych informacji. W ten sposób system przestaje „widzieć" rozpraszające sygnały (np. zachętę do wspólnej zabawy wysyłaną przez członków stada) i kontynuuje bieżące zadanie aż do momentu, kiedy będzie ono zrealizowane (a przynajmniej znacznie zaawansowane) albo do chwili, kiedy pojawi się informacja, której zignorowanie przez agenta byłoby niekorzystne np. stresor lub wysoko ceniona nagroda (na diagramie filtracja przeprowadzana przez podsystem P-MM wyrażona jest poprzez zamianę symboli $[o\ldots]_{t+1}$ na o_{t+1}, oraz $[r_{ś}\ldots]_{t+1}$, $[r_{h}\ldots]_{t+1}$ na r_{t+1}; w ten sposób obserwacje i nagrody wejściowe uzyskują status obserwacji i nagród istotnych dla realizowanego celu). Warto przypomnieć, że tego typu filtr jest bezpośrednią konsekwencją hipotezy bramkowania dopaminowego, zaproponowanej przez O'Reilly'a, Bravera i Cohena (O'Reilly i in., 1999). Godny podkreślenia jest fakt, że wskazana „bramka" posiada również zdolność uczenia się. Implementacją neuronalną tego mechanizmu są struktury hipokampu, które umożliwiają agentowi zapamiętywanie reprezentacji (obserwacji) uznanych za ważne z perspektywy realizowanego celu. Istotne są te reprezentacje, które pojawiły się w trakcie realizacji celu oraz którym towarzyszył niezerowy błąd predykcji nagrody, sygnalizujący informację wartościującą, której nie spodziewał się podsystem P-RL. W badaniach eksperymentalnych są to na przykład określone bodźce warunkowe, np. kroki opiekuna sygnalizujące porę karmienia. Zdaniem Montague, tego typu błąd predykcji nagrody wzmacnia lub osłabia, podczas realizowania kolejnych epizodów, określone związki między obserwacjami a nagrodami. W przyszłości tego typu asocjacja pozwoli filtrowi funkcjonującemu w P-MM zareagować, gdy w

strumieniu napływających obserwacji pojawi się zapamiętana reprezentacja skorelowana z określoną nagrodą.

Poza stabilizacją podsystemu P-RL, podsystem P-MM realizuje jeszcze jedną bardzo ważną funkcję – decyduje o zmianie celu. Agent musi co jakiś czas sprawdzać czy w środowisku lub wewnątrz organizmu nie zaszły istotne zmiany, z powodu których należałoby dezaktywować bieżący cel i wybrać nowy (c_x zmienia się w c_y). Tego typu dyspozycja wymaga od podsystemu P-MM **zdolności do rozpoznawania** informacji w strumieniu napływających obserwacji i nagród, które „zwiastują" ważne z perspektywy organizmu możliwości, np. pozyskanie cennych nagród niezwiązanych z aktualnie realizowanym celem lub rozpoznanie niedoboru energetycznego doświadczanego jako głód. Jeśli w tego rodzaju strumieniu rozpoznana zostanie informacja wymuszająca wstrzymanie realizacji bieżącego celu, wówczas P-MM uruchomi odpowiedni proces wymiany celu na nowy (c_x / c_y).

Przypomnijmy, że w podsystemie P-RL cel to nic innego, jak zbiór następujących reprezentacji:

1. reprezentacji nagród (r) charakterystycznych dla danego celu,
2. reprezentacji stanów świata (s) prowadzących do pozyskania nagród (warto zauważyć, że stany świata (s) są silnie skorelowane z obserwacjami (o) dostarczanymi przez podsystem sensoryczny),
3. reprezentacji zachowań (z) niezbędnych do pozyskania nagród,
4. reprezentacji funkcji wartości V wyznaczającej dla danego stanu świata (s) jego użyteczność z perspektywy realizowanej polityki doboru zachowań π;
5. hiper-parametrów[111] procesu uczenia się takich jak: wartość dyskonta (γ), tempo uczenia się (β), stopień eksploracji (ε).

[111] „Many models have important parameters which cannot be directly estimated from the data. For example, in the K-nearest neighbor classification model, a new sample is predicted based on the K-closest data points in the training set. […]. This type of model parameter is referred to as a tuning parameter because there is no analytical formula available to calculate an appropriate value." (Kuhn, 2013, s. 64–65). Hiperparametr to parametr, który jest ustawiany przed rozpoczęciem procesu uczenia. Parametry te są dostosowywane do specyfiki problemu i mogą bezpośrednio wpływać na to, jak dobrze model się uczy (*Hyperparameter*, 2019).

Powyższy zbiór reprezentacji, składających się na definicję celu w metodzie uczenia się ze wzmacnianiem, musi być przechowywany w pamięci długotrwałej[112]. Dzieje się tak dlatego, że organizmy biologiczne nie mogą – ze względu na potrzeby energetyczne – pozwolić sobie na uczenie się i optymalizowanie danego celu bez przerwy. Proces ten musi być co jakiś czas przerywany (np. na zaspokojenie głodu, odpoczynek, itp.), a następnie wznawiany w sprzyjających okolicznościach. Znaczy to, że wartości poszczególnych reprezentacji (1-5), konstytuujących dany cel, muszą być co jakiś czas „odtwarzane", by proces uczenia się optymalnej strategii doboru zachowań mógł być kontynuowany. Równocześnie, w trakcie realizacji celu, podsystem P-RL powinien posiadać możliwość „odczytu z" oraz „zapisu do" pamięci następujących typów reprezentacji: funkcji wartości V dla poszczególnych stanów świata, bieżących wartości parametrów kalibrujących proces uczenia się oraz aktualnego stanu strategii doboru zachowań π. Postać strategii zmieniać się zatem będzie z każdą interakcją między agentem a środowiskiem, niezależnie od tego czy realizacja celu zostanie przerwana, czy będzie po pewnym czasie wznowiona. Znaczy to, że proces uczenia się agenta składa się z różnych epizodów zależnych od jego bieżących potrzeb.

W kontekście powyższych uwag należy sformułować pytanie o to, czy rozszerzenia przedstawione w Modelu 1.1 są wystarczające, by wyjaśnić złożone działanie intencjonalne? Odpowiedź na to pytanie, podobnie jak w przypadku odpowiedzi na podobne pytanie odnoszące się do prostszego Modelu 1.0, jest negatywna. Niewątpliwie, dodanie do struktury Modelu 1.0 podsystemów motywacji i monitorowania (P-MM) oraz homeostazy (P-H) umożliwia przezwyciężenie ograniczeń „jedno-celowego" Modelu 1.0, charakterystycznego dla agentów sztucznych (np. robotów) oraz prostych organizmów. Zdolność agenta do przełączania się pomiędzy różnymi celami umożliwia mu utrzymanie homeostazy oraz reakcję na zmieniające się warunki otoczenia, przy czym zakres jego adaptacyjności nadal jest mocno ograniczony. Głównym powodem takiego stanu rzeczy jest następująca własność systemu zgodnego z Modelem 1.1: tego typu agenci dysponują nielicznym i zamkniętym zbiorem reakcji na zachodzące zmiany, gdyż nie posiadają mechanizmów, które pozwalałyby im wykroczyć poza wrodzone reprezentacje nagród (r),

[112] Zaproponowany model zakłada, że każdy z podsystemów dysponuje pamięcią długotrwałą, w której przechowuje specyficzne dla siebie reprezentacje. Takie ujęcie nie wyklucza, że tego typu pamięć może być współdzielona pomiędzy podsystemami lub być zrealizowana jako niezależny podsystem, centralizujący tego typu funkcję. Przyjęto, że rozstrzygnięcie tego typu kwestii nie wpływa istotnie na ogólność i trafność modelu.

zachowań (z) oraz obserwacji (o). Wszystkie wymienione elementy są w dużym zakresie zdeterminowane genetycznie i mogą się zmienić w efekcie pojawienia się stosownej mutacji genów. Znaczy to, że jeśli w środowisku zabraknie określonego rodzaju nagród lub spadnie ich dostępność (np. ze względu na wzrost populacji), to istnieje duże ryzyko, że organizm zginie. Innymi słowy, w tego typu przypadkach możliwości adaptacyjne mają znacznie mniejszy zakres i są ściśle skorelowane z cechami zajmowanej przez dany gatunek niszy ekologicznej. Zwierzęta wyższego rzędu (ptaki, ssaki), u których – w zależności od ich wieku – obserwuje się znaczące różnice w formie oraz złożoności zachowań, najprawdopodobniej dysponują mechanizmem, który pozwala im powoływać do życia nowe typy reprezentacji i modyfikować podejmowane działania. Proponuję, mając na uwadze wyżej opisane zależności, wprowadzić do Modelu 1.1 dwa uzupełnienia: (1) wzbogacenie wybranych podsystemów o zdolność do tworzenia nowych reprezentacji oraz (2) dołączenie mechanizmu przypisującego status nagrody stanom intencjonalnym, tzn. traktującego przekonania, pragnienia, intencje i pokrewne stany umysłowe jako mające tak doniosłą wartość dla agenta, że może ona decydować o podjęciu danego działania intencjonalnego (np. pragnienie zwycięstwa skłaniające do kandydowania w wyborach prezydenckich).

5.3.3 Model 1.2 - działanie intencjonalne z kreatorem nowych typów zachowań oraz ewaluatorem stanów umysłowych jako nowego typu nagród

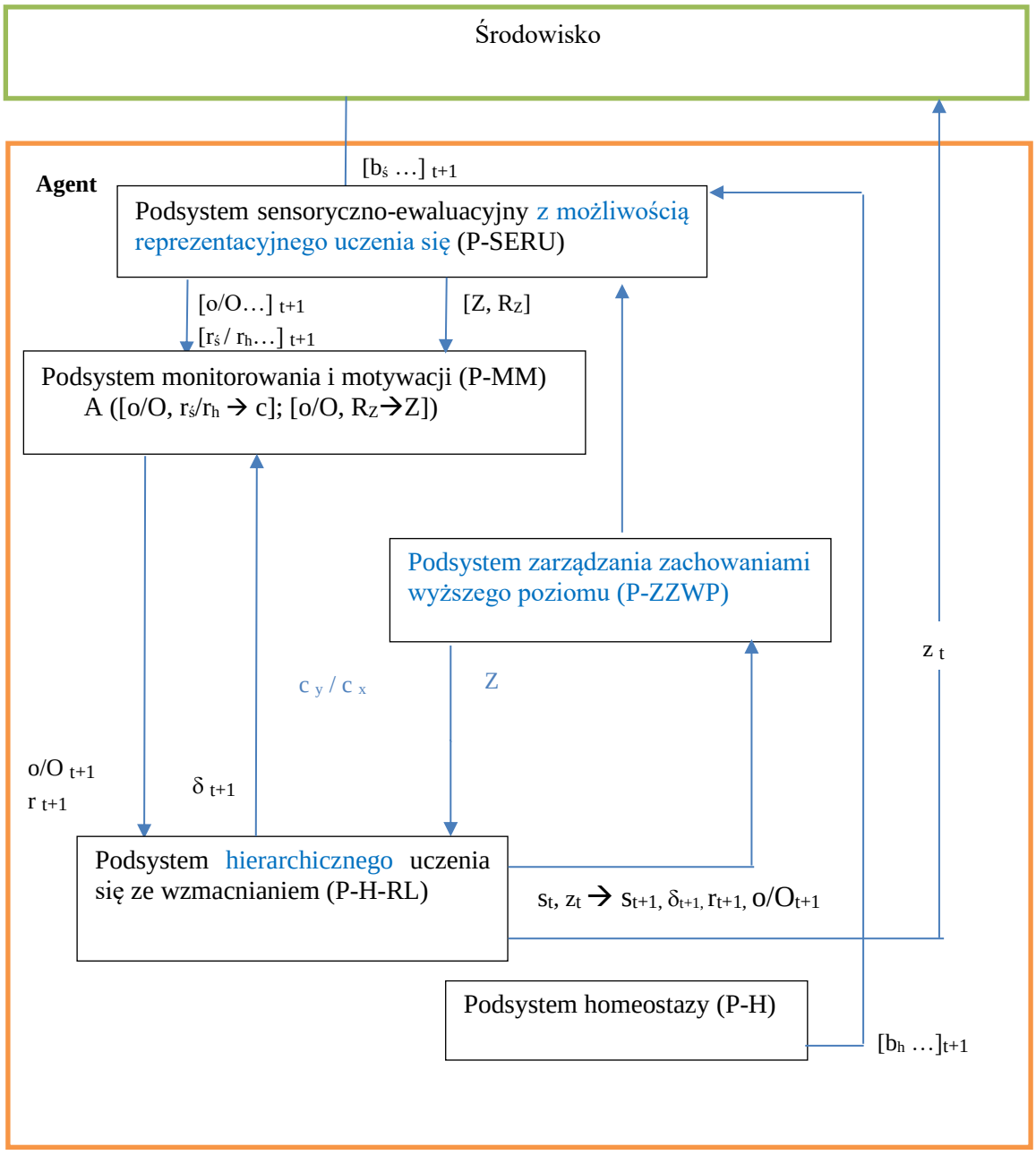


Diagram 12. Model 1.2 działania intencjonalnego z kreatorem nowych typów zachowań oraz ewaluatorem stanów umysłowych jako nowego typu nagród.

Legenda symboli (uzupełnienie w porównaniu z Modelem 1.1):

- O_{t+1} – reprezentacja stanu środowiska w chwili t+1, utworzona przez podsystem sensoryczno-ewaluacyjny dysponujący zdolnością uczenia się nowych typów

reprezentacji; obserwacja typu ‘O’ – w odróżnieniu od obserwacji wrodzonych ‘o’ – ma charakter dynamiczny i zmienia się w czasie – wraz z dojrzewaniem organizmu;

- A ([o/O, $r_ś/r_h$/R → c], [o/O, R → Z]) – asocjacje łączące obserwacje, które dotyczą środowiska z nagrodami oraz celami (szczególnym przypadkiem celu może być wysokopoziomowe zachowanie Z – więcej na ten temat w dalszej części rozdziału); wyróżnione w podsystemie P-MM asocjacje umożliwiają: (1) aktywowanie celów ‘c’ na podstawie określonych obserwacji ‘o/O’ oraz związanych z nimi nagród lub (2) stabilizowanie celów poprzez filtrację obserwacji i nagród nieistotnych z perspektywy ich realizacji;
- Z – reprezentacja zachowania wyższego poziomu (patrz: koncepcja opcji zaprezentowana w rozdziale drugim w sekcji: *Hierarchiczne uczenie się ze wzmacnianiem*) utworzona przez podsystem zarządzający zachowaniami wyższego poziomu (P-ZZWP) w trakcie rozwoju ontogenetycznego; niskopoziomowe reprezentacje zachowań (z_t), skorelowane z nimi reprezentacje stanów świata (s_t) oraz obserwacje (O/o_t) – wraz z relacjami istniejącymi między nimi (patrz: korelacja „s_t, z_t → s_{t+1}, δ_{t+1}”) – umożliwiają utworzenie tego typu reprezentacji (Z);
- R_Z – reprezentacja określająca wartość zachowania wyższego poziomu (Z), wyznaczona przez moduł ewaluujący podsystemu P-SERU; za pomocą tego typu reprezentacji zachowanie Z zaczyna być traktowane jak nagroda, która wpływa na dobór zachowań w podsystemie P-H-RL.

Uzasadnienie

System kontroli działań, jak zostało to przedstawione na powyższym diagramie, wzbogacony został o podsystem zarządzania zachowaniami wyższego poziomu (P-ZZWP). Jego główne zadanie polega na wytwarzaniu nowych typów zachowań (Z) przy wykorzystaniu reprezentacji niskopoziomowych stosowanych przez pozostałe podsystemy. W ten sposób zachowania wyższego poziomu obejmują złożone sekwencje zachowań prostszych. W obszarze uczenia maszynowego o algorytmach posiadających zdolność organizowania zachowań w hierarchię mówi się, że wprowadzają one do działań czasową abstrakcję (*temporal abstraction*) (Sutton i in., 1999). Podobnie, podsystem sensoryczny uzyskał zdolność wytwarzania obserwacji (O), które cechują się większą uniwersalnością, czyli pozwalają agentowi trafnie rozpoznać dany stan środowiska,

pomimo pojawiających się zniekształceń i zaburzeń w odbiorze sygnałów w nim funkcjonujących. Innymi słowy, podsystem sensoryczno-ewaluacyjny (P-SE) uzupełniony został o zdolność uczenia się nowych reprezentacji (stąd zmiana nazwy podsystemu na P-SERU). W konsekwencji, podsystem P-RL został poszerzony o możliwość stosowania zachowań wyższego rzędu (Z). Tego typu reprezentacje mogą pełnić we wskazanym podsystemie dwojaką rolę: (1) mogą pełnić funkcję abstrakcyjnych umiejętności optymalizujących proces pozyskiwania nagród albo (2) mogą pełnić funkcję nagradzającego aktu behawioralnego, którego realizacja jest dla agenta nagrodą samą w sobie – stąd nazwa podsystemu. W rezultacie, jego charakterystyka została rozszerzona do postaci podsystemu hierarchicznego uczenia się ze wzmacnianiem (P-H-RL). Aby nowe reprezentacje sensoryczne (O) były w pełni użyteczne, odpowiednie rozszerzenia powinny objąć podsystem motywacji i monitorowania, na wejściu którego – oprócz typów obserwacji (o) rozpoznawanych od urodzenia – mogą pojawić się również reprezentacje nabyte (O) oraz nagrody (R_Z), które są związane z zachowaniami wyższego poziomu. Wymienione zmiany łącznie zapewniają efektywniejszą i elastyczniejszą kontrolę zachowań, niż w przypadku systemu opisanego w modelu 1.1, który funkcjonuje wyłącznie na podstawie reprezentacji wrodzonych wykorzystujących bardzo wąski wycinek informacji pochodzących ze środowiska. Istnienie mechanizmów, które mogą „produkować" nowe typy reprezentacji, potwierdzone jest licznymi obserwacjami behawioralnymi, w szczególności tymi, które pokazują różnice pomiędzy osobnikami młodymi a dorosłymi (Sadowski, 2012)). Efekt ten można wyjaśnić, odnosząc się do szeroko pojętego mechanizmu uczenia się, który obejmuje zarówno reprezentacje wspierające proces pozyskiwania nagród, jak i proces ich definiowania. Każdy nowo zdefiniowany typ nagrody włączony w podsystem P-H-RL, zgodnie z zasadą działania algorytmu uczenia się ze wzmacnianiem, będzie prowadził do pojawienia się specyficznych form zachowań. Warto przypomnieć, że pojęcie nagrody w metodzie uczenia się ze wzmacnianiem ma charakter abstrakcyjny i może odnosić się do obiektów różnych typów. Podsystem P-H-RL reaguje na trojakiego rodzaju nagrody:

- obiekty dostępne w środowisku (np. nowy rodzaj pokarmu (nagroda dodatnia) lub trucizna (nagroda ujemna)),
- akty behawioralne (np. wokalizacje odstraszające drapieżniki (nagroda dodatnia) lub zachowania agresywne powodujące wykluczenie z grupy społecznej (nagroda ujemna)),

- stany wewnętrzne organizmu (np. pobudzenie będące skutkiem wysokiego poziomu adrenaliny (nagroda dodatnia) lub stany apatii będące skutkiem długotrwałego stresu (nagroda ujemna)).

Nowe typy reprezentacji, jak już wspomniałem, umożliwiają agentowi efektywniejsze eksplorowanie środowiska, a poprzez to – skuteczniejsze osiąganie zarówno wrodzonych, jak i nabytych celów. Obserwacje typu O pozwalają podsystemowi sensoryczno-ewaluacyjnemu prawidłowo identyfikować obiekty lub stany środowiska na podstawie niepełnych i często zaburzonych bodźców (patrz: problem kontekstowej niezmienności (*contextual invariance*) (Friston, 2003)). Natomiast reprezentacje zachowań wyższego poziomu (Z), funkcjonujące jako abstrakcyjne umiejętności umożliwiają agentowi, wykorzystującemu wcześniejsze doświadczenia, znacznie skrócić czas eksploracji środowiska. O ile wymienione typy reprezentacji wspierają głównie mechanizm kontroli zachowań, o tyle nowe typy nagród 'R' przyczyniają się do pojawienia się zupełnie nowych form zachowań (patrz: nagradzające akty behawioralne oparte na nagrodach typu R_Z, np. tresura zwierząt). Warto w tym miejscu zauważyć, że „zyski" behawioralne osiągnięte na skutek dysponowania tego typu reprezentacjami zdobywane są kosztem wydatkowania odpowiedniej ilości energii, a zatem ich pojawienie się nie zawsze musi być korzystne dla organizmu. W związku z tym należy oczekiwać, oprócz możliwości tworzenia nowych typów reprezentacji w układzie nerwowym, że będą istniały mechanizmy, które z czasem usuną lub zmodyfikują niektóre reprezentacje do bardziej użytecznej postaci. Mamy tu zatem do czynienia ze swoistym cyklem życia, który obejmuje (a) utworzenie, (b) modyfikację albo (c) usunięcie reprezentacji z systemu[113] (oczywiście, w trakcie trwania cyklu realizowane są odczyty i aktywacje nagród prowadzące do określonych działań). Występowanie takiego cyklu oznacza, że pewne formy zachowań będą się pojawiały i z czasem zanikały. Pozwala też zaobserwować zasadniczą różnicę w reakcjach na nagrody wrodzone (patrz: koncepcja nagród podstawowych w ujęciu Reada Montague'a (*primary rewards*) (Montague, 2006, s. 128)) oraz nabyte. O ile te ostatnie mogą w pewnym momencie zaniknąć (zrealizować pełen cykl życia), o tyle nagrody podstawowe nigdy nie mogą zostać usunięte z systemu, co najwyżej chęć ich pozyskania może zostać

[113] Warto zauważyć, że – zgodnie z analizami Rolfa Landauera oraz Charlesa Bennetta – poszczególne typy operacji (zapis, modyfikacja, usuwanie), które realizowane są na poziomie mózgu traktowanego jako system przetwarzający informacje, różnią się pod względem kosztów energetycznych niezbędnych do ich przeprowadzenia (Montague, 2006, s. 66). Znaczy to, że dopóki organizmowi nie grozi „kryzys zasobowy", dopóty operacje usuwania zbędnych reprezentacji mogą być odraczane w czasie.

wyhamowana – i to na stosunkowo krótki czas. Ten szczególny status nagród podstawowych jest ważnym ewolucyjnym zabezpieczeniem, które chroni organizm przed przewartościowaniem nagród nabytych. Gatunki, które potrafią w jakimś stopniu oprzeć się „dyktatowi" nagród biologicznych, mają znacząco większe możliwości dostosowawcze do zmieniających się warunków otoczenia, gdyż mogą poświęcić część zasobów na zdobywanie nowych typów nagród, w szczególności związanych z funkcjonowaniem agenta w grupie społecznej (patrz: zachowania protokulturowe[114]). Szczególnie widoczne jest to w przypadku gatunku ludzkiego, który w niezwykłym wręcz zakresie potrafi wyhamowywać impulsy biologiczne, a w ekstremalnych przypadkach – zawiesić działanie instynktu przetrwania (patrz: historia członków sekty *Heaven's Gate* (Montague, 2006, s. 88)). Tego typu dyspozycja wymaga posiadania odpowiednio elastycznego podsystemu monitorowania i motywacji, który przynajmniej na jakiś czas potrafi „zawiesić" wpływ nagród podstawowych, oraz odpowiednio złożonego podsystemu zarządzania reprezentacjami, zdolnego do realizowania nowych celów.

Ważnym etapem w ewolucji złożonych form zachowań był okres, w którym istotną rolę zaczęły odgrywać działania niezwiązane z bezpośrednim zaspokajaniem podstawowych potrzeb. Zwierzęta zaczęły realizować m.in. następujące cele: zdobycie określonego miejsca w grupie (np. walka o pozycję samca alfa), przyciągnięcie uwagi osobnika płci przeciwnej (np. rytuały godowe[115]) czy komunikacja w celu zaawansowanego, wymagającego kooperacji, eksplorowania lub monitorowania otoczenia. Wymienione tu przypadki są przykładami działań, które nabierają znaczenia ze względu na interakcje występujące wewnątrz grupy. Wydawać by się mogło, że tego typu zachowaniami rządzą zupełnie inne mechanizmy niż te, które dotyczą zaspokojenia podstawowych potrzeb. Jednak z ewolucyjnego punktu widzenia bardziej prawdopodobna wydaje się hipoteza mówiąca, że formy zachowania, które są złożone i podlegające złożonym mechanizmom kontroli, wyewoluowały jako nadbudowa i uzupełnienie mechanizmu bazowego, tj. uczenia się ze wzmacnianiem.

Powstaje pytanie: jak to się dzieje, że określone akty behawioralne (np. zachowania społeczne zwierząt (Korpikiewicz, 2017)), które nie są zdeterminowane genetycznie, stają

[114] Istotą protokultury jest to, że wiedza nabyta indywidualnie zostaje zachowana także po śmierci osobnika, który ją pierwotnie nabył („dziedziczenie" kultury) (McGrew, 2003).

[115] Wskazany przykład ma charakter poglądowy, gdyż w wielu przypadkach rytuały godowe są wrodzone. (patrz: zachowania godowe ptaków oraz efekt tzw. wpajania (*imprinting*)) (Sadowski, 2012, s. 468).

się celami samymi w sobie? Odpowiedź na to pytanie odsyła do mechanizmów odpowiedzialnych za wytwarzanie nowych typów reprezentacji, a w szczególności – nowych typów nagród. Niestety, badacze, których prace poddano analizie w rozdziałach trzecim i czwartym niniejszej rozprawy, nie uwzględniali podniesionych wyżej kwestii: (a) przebiegu procesów nabywania nowych typów obserwacji (O), (b) zachowań wyższego poziomu (Z) oraz (c) nagród (R_Z) powiązanych z zachowaniami wyższego poziomu. Wskazać można koncepcje i hipotezy, które mogłyby pomóc w dookreśleniu natury tego typu procesów, problem polega na tym, że do tej pory nie były one uwzględniane ani przez psychologów intencji, ani przez badaczy zajmujących się hipotezą dopaminergicznego błędu predykcji nagrody. Z perspektywy proponowanego tu zintegrowanego modelu złożonych działań intencjonalnych jest to istotna luka, którą należałoby wypełnić. Proponuję, by uzupełnić model i zapełnić wskazaną lukę, przez przyjęcie następujących hipotez odnoszących się do poszczególnych typów reprezentacji:

- (H1) Obserwacje (O) są efektem kodowania predykcyjnego (*predictive coding*), implementującego hipotezę mózgu bayesowskiego (*learning based on empirical Bayes*) (Friston, 2003),
- (H2) Zachowanie wyższego poziomu (Z) jest sekwencją zachowań niskopoziomowych, określoną w ramach abstrakcyjnego modelu środowiska, umożliwiającą agentowi przemieszczanie się pomiędzy jego kluczowymi stanami (Lakshminarayanan i in., 2016; Singh i in., 2005; Yao i in., 2014),
- (H3) Nagrody reprezentujące zachowania wysokiego poziomu (R_Z) są efektem działania uniwersalnego mechanizmu szacowania wartości nagrody, który działa w obrębie podsystemu sensoryczno-ewaluującego (patrz: hipoteza wspólnej neuronalnej waluty) (Levy & Glimcher, 2012).

Obserwacje (O) jako efekt kodowania predykcyjnego – (H1)

Niewątpliwie najbardziej wiarygodna i najsolidniej opracowana jest hipoteza (H1) dotycząca kodowania predykcyjnego odnoszącego się do danych sensorycznych. Jej współautorem jest Karl Friston, czołowy przedstawiciel tzw. teoretycznej neurobiologii (*theoretical neurobiology*), w ramach której do analizy danych empirycznych pozyskiwanych za pomocą fMRI stosuje się zaawansowane modele obliczeniowe. Wśród wielu dokonań Fristona ważne miejsce zajmuje model uczenia się reprezentacji sensorycznych (*representational learning*) niezależnych od kontekstu. Koncepcja, którą

proponuje angielski neuronaukowiec, opiera się na następującym spostrzeżeniu: związki przyczynowe między zmysłami a otoczeniem wywołują wrażenia zmysłowe (*sensory input*), które są silnie zależne od kontekstu. Obiekt widziany w spoczynku wywołuje zupełnie inne wrażenia, niż ma to miejsce w przypadku tego samego obiektu widzianego w ruchu. Dodatkowo, różnego rodzaju interakcje między obiektami powodują, że do układu sensorycznego docierają bodźce odpowiadające złożonym kombinacjom związków przyczynowych. Znaczy to, że wiele różnych przyczyn może powodować tę samą odpowiedź układu sensorycznego (jest to więc odwzorowanie wielo-jednoznaczne; np. podobny kształt posiada trzonek do grabi, miotła, kula dla osób niepełnosprawnych, drążek do podciągania się, itp.), zarazem jednak wiele różnych stanów układu sensorycznego może być wywołanych przez tę samą przyczynę (odwzorowanie jedno-wieloznaczne; np. odmienne kształty widzianego z różnych ujęć zegarka pochodzą od tego samego przedmiotu). Problem identyfikacji źródła reprezentacji sensorycznej jest więc niezwykle złożony, gdyż rozpoznanie przyczyn na podstawie niejednoznacznych bodźców, które na dodatek uwikłane są w różne konteksty[116], wymaga stosowania różnorodnych strategii interpretacyjnych. Wskazana niejednoznaczność sprawia, zdaniem Fristona, że problem ustalenia **przyczyn** wrażeń sensorycznych jest z zasady niedookreślony (*under-determined*) lub źle określony (*ill posed*).

Friston zapisuje relację między przyczynami (stanami środowiska oddziałującymi na organizm) a ich skutkami, czyli wrażeniami sensorycznymi w postaci deterministycznej nieliniowej funkcji generatywnej (*deterministic non-linear generative function*) (Friston, 2003, s. 1331):

$$u = G(v,\theta),$$

gdzie v to wektor przyczyn pochodzących ze środowiska, a u to odpowiadające im dane sensoryczne. $G(v,\theta)$ – to funkcja generująca dane sensoryczne na podstawie wektora przyczyn v oraz odpowiedniego modelu „skalibrowanego" za pomocą zbioru parametrów θ[117]. Nieliniowość obecna w definicji funkcji G odnosi się do sytuacji, w której następuje interakcja pomiędzy przyczynami wchodzącymi w skład wektora v, np. kiedy ruch obiektu wpływa na odbicie światła i w konsekwencji na cechy oraz na intensywność bodźców

[116] „At a more cognitive level the cause associated with the word 'HAMMER' will depend on the semantic context (that determines whether the word is a verb or a noun)." (Friston, 2003, s. 1331).

[117] „We shall see later that the parameters correspond to connection strengths in the brain's model of how inputs are caused." (Friston, 2003, s. 1331).

docierających do aparatu percepcyjnego. W takim przypadku bodźce są wypadkową wielu przyczyn, co jeszcze bardziej komplikuje ich rozpoznanie.

Z powyższego wynika, że problem rozpoznawania obiektów w środowisku można sprowadzić do problemu znalezienia funkcji odwrotnej do funkcji *G*. Z jej pomocą możliwe staje się zidentyfikowanie przyczyn ν na podstawie danych sensorycznych *u*. Główny problem polega na tym, że funkcja $G(\nu,\theta)$ w określonych przypadkach jest nieodwracalna, tzn. nie da się w prosty sposób na podstawie *u* określić, które przyczyny ν wygenerowały wrażenia (np. dotykowe). Brak możliwości analitycznego wyznaczenia funkcji odwrotnej przejawia się również w tym, że probabilistyczne metody uczenia maszynowego stosowane w takich przypadkach prowadzą do eksplozji kombinatorycznej[118]. Pomimo tych ograniczeń, to właśnie stochastyczne modele są obecnie najbardziej popularnymi metodami wyjaśniającymi procesy rozpoznawania reprezentacji przyczyn. W podejściu tym problem odwracalności deterministycznej nieliniowej funkcji *G* zastępuje się problemem możliwości parametryzacji funkcji gęstości odwrotnego prawdopodobieństwa warunkowego (*existence of an inverse conditional probability (i.e. recognition) density that can be parameterized*). Wymagana jest współpraca dwóch rodzajów modeli, aby tak postawiony problem rozwiązać: modeli rozpoznawczych (*recognition*) oraz generatywnych (*generative*). Pierwszemu rodzajowi odpowiada funkcja, która na podstawie danych sensorycznych wyznacza leżące u ich podstaw przyczyny (Roz: DS→P)[119], natomiast drugiemu rodzajowi odpowiada funkcja, która na podstawie przyczyn wyznacza dane sensoryczne (Gen: P→DS)[120]. Zdaniem Fristona, odpowiednie połączenie wskazanych modeli, w ich probabilistycznej, a nie deterministycznej wersji, pozwala efektywnie rozwiązać problem nieodwracalności funkcji *G*. Składowe poszczególnych modeli można scharakteryzować na odpowiednim poziomie ogólności w sposób opisany poniżej.

Model generatywny składa się z: (1) rozkładu prawdopodobieństwa typu prior dla wektora przyczyn $\nu - p(\nu;\theta)$ oraz (2) wiarygodności definiowanej jako prawdopodobieństwo uzyskania danych sensorycznych (*u*) pod warunkiem zaistnienia wektora przyczyn ν - $p(u|\nu;\theta)$. Możliwe jest wyznaczenie dla danego modelu

[118]Eksplozja kombinatoryczna powoduje, że liczba sposobów, na jakie stochastyczne modele generatywne mogą wytworzyć dany wzorzec, rośnie wykładniczo wraz z jego długością.

[119] Na poziomie mózgu funkcji tej odpowiadają połączenia wyprzedzające (*forward*).

[120] Na poziomie mózgu funkcji tej odpowiadają połączenia wsteczne (*backward*).

generatywnego tzw. rozkładu brzegowego *p(u; θ)* (*marginal distribution*) na podstawie powyższych składowych oraz w wyniku procesu uczenia się. Model ten z czasem staje się dobrym przybliżeniem *p(u),* czyli faktycznym rozkładem prawdopodobieństwa danych sensorycznych.

Model rozpoznawczy[121], złożony z funkcji gęstości prawdopodobieństwa *p(u; θ)*, prawdopodobieństwa prior *p(ν;θ)* oraz wiarygodności *p(u|ν;θ)*, pozwala wyznaczyć *p(ν|u;θ)*, czyli prawdopodobieństwo *a posteriori* określające najbardziej prawdopodobną przyczynę powstania danych sensorycznych. Obecnie dysponujemy całym szeregiem metod implementujących proces wyznaczania *p(ν|u;θ)*. Kiedy jednak pod uwagę weźmie się wymagania neurobiologiczne ustalone na podstawie architektury mózgu, to najbardziej realistycznym podejściem okazuje się połączenie dwóch metod: (1) kodowania predykcyjnego oraz (2) hierarchiczno-empirycznego uczenia bayesowskiego. Friston twierdzi, że kombinacja wymienionych metod pozwala rozwiązać trudny problem oszacowania wartości prawdopodobieństw typu prior (tego typu wymaganie spełnia empiryczne uczenie bayesowskie), a także problem nieodwracalności funkcji *G* (ten element obsługuje kodowanie predykcyjne wraz z mechanizmem minimalizacji błędu predykcji).

Zarysowana powyżej Fristonowska koncepcja nabywania reprezentacji sensorycznych jest obecnie przedmiotem licznych analiz oraz badań empirycznych. Wielu czołowych badaczy zajmujących się modelowaniem funkcji mózgu (m.in. Anil Seth, Rafał Bogacz) postrzega kodowanie predykcyjne oraz mechanizm redukcji błędu predykcji przy pomocy metod optymalizujących wolną energię (*free energy*) jako jedną z najważniejszych zasad organizujących jego działanie (patrz: predykcyjna teoria umysłu oraz hipoteza mózgu bayesowskiego (Bogacz, 2017; Millidge i in., 2021)). W efekcie, w wielu obszarach badawczych podejmowane są próby zastosowania wskazanego modelu do opisu zjawisk niezwiązanych bezpośrednio z przetwarzaniem danych sensorycznych (np. dotyczących organizacji zachowań)[122]. Trudno obecnie ocenić, w jakim stopniu takie rozszerzenia okażą się poznawczo płodne. Wydaje się jednak, iż opisany wyżej mechanizm uczenia się

[121] $p(\nu|u;\theta) = [p(u|\nu;\theta) * p(n;\theta)] / p(u;\theta)$.

[122] Dobrym przykładem może być praca Lisy Feldman Barret dotycząca konstruktywistycznej natury emocji. W artykule *The theory of constructed emotion: An active inference account of interoception and categorization* badaczka ta, powołując się na prace m.in. Fristona, w zasadniczy sposób kwestionuje dotychczasowe podejście dotyczące funkcji emocji (Barrett, 2016). Por. także ujęcie popularniejsze w Barrett (2020).

reprezentacji i rozpoznawania obiektów trafnie wyjaśnia przebieg przetwarzania informacji percepcyjnej, co – z perspektywy prowadzonych w niniejszej pracy rozważań – jest ważnym argumentem na rzecz wyodrębnienia podsystemu sensoryczno-ewaluacyjnego z mechanizmem reprezentacyjnego uczenia się (P-SERU) (patrz: Rys. 3 przedstawiający Model 1.2). Organizmy, które dysponują tego typu rozszerzeniem, niewątpliwie zwiększają zakres postrzeganych przez siebie obiektów i w ten sposób wzmacniają skuteczność działania[123].

Zachowanie wyższego poziomu (Z) jako sekwencja zachowań z poziomu niższego – (H2)

Zaproponowana przez Fristona koncepcja nabywania reprezentacji sensorycznych jest przykładem dobrze ugruntowanej od strony teoretycznej i empirycznej hipotezy badawczej. Niestety, w przypadku zachowań wyższego poziomu (Z) nie dysponujemy tak zaawansowaną i dojrzałą koncepcją, jak kodowanie predykcyjne oparte na hierarchiczno-empirycznym uczeniu bayesowskim. Przekonanie, iż zachowania wyższego poziomu są niezbędne dla sprawnego funkcjonowania organizmów biologicznych oraz odpowiednio zaawansowanych robotów, jest powszechnie podzielane przez badaczy zajmujących się uczeniem maszynowym (Sutton i in., 1999). Przedstawiona w rozdziale trzecim koncepcja hierarchicznego uczenia się ze wzmacnianiem pokazuje, jak zachowania wysokiego poziomu (tzw. opcje) wpływają na działanie algorytmu RL, jednakże nie znajdujemy w niej wyjaśnień odnoszących się do sposobu, w jaki tego typu hierarchizacja jest nabywana przez system.

Ponad dwadzieścia lat minęło od publikacji artykułu Suttona i współpracowników. W tym okresie pojawiło się kilka oryginalnych koncepcji rozwijających teorię zachowań wyższego poziomu w algorytmach uczenia się ze wzmacnianiem. Wśród dostępnych propozycji znaleźć można opracowaną przez Suttona i zespół koncepcję „modelu uniwersalnej opcji" (*universal option model*) (Yao i in., 2014), podejście „hierarchicznego głębokiego uczenia się ze wzmacnianiem: integrującego czasowe abstrahowanie z wewnętrzną motywacją" (*hierarchical deep reinforcement learning: integrating temporal abstraction and intrinsic motivation*) (Kulkarni i in., 2016), które zaproponował zespół z

[123] Równocześnie należy pamiętać, że proces uczenia się nowych reprezentacji sensorycznych wiąże się z określonym wydatkiem energetycznym, który nie musi być opłacalny, dlatego ważna jest równowaga pomiędzy procesami reprezentacjo-twórczymi a kosztami energetycznymi poniesionymi w związku z ich obsługą.

MIT czy wreszcie rozwiązanie przygotowane przez grupę badaczy z Politechniki w Madrasie oparte na odkrywaniu opcji wpisanych w przestrzenno-czasowe klastrowanie (*option discovery in hierarchical reinforcement learning using spatio-temporal clustering*) (Lakshminarayanan i in., 2016).

Można powiedzieć, że w wymienionych koncepcjach dąży się do „oderwania" definicji opcji od kontekstu, w którym została ona opracowana. Celem takiego „oderwania" jest uniwersalizacja doświadczeń zdobytych w danym obszarze problemowym i przeniesienie ich na inne typy problemów.

Nawet w przypadku złożonych scenariuszy gier (np. *Montezuma's Revenge*), jak pokazują eksperymenty z wykorzystaniem emulatora Atari, sztuczny agent korzystający z tego typu metod jest w stanie w istotny sposób skrócić proces uczenia się i osiągać wyniki znacznie lepsze od wyników systemów opartych wyłącznie na podstawowej wersji algorytmu uczenia się ze wzmacnianiem. Trudno obecnie ocenić, które z wymienionych powyżej podejść uzyska w przyszłości status wiarygodnego biologicznie modelu złożonych zachowań. Być może będą to rozwiązania inspirowane badaniami neurobiologicznymi. Na przykład, odpowiednikiem próbkowania trajektorii zachowań w celu utworzenia przybliżonego modelu środowiska (*sample trajectories to construct an approximate estimate [of environment])* (patrz: rozwiązanie zespołu z Politechniki w Madrasie, (Lakshminarayanan i in., 2016)) może być podproces konsolidacji śladów pamięciowych realizowany przez powtórzenia podczas snu przeżyć zrealizowanych wcześniej na jawie (*experience replay*) (O'Neill i in., 2010). Są to tylko wstępne przypuszczenia, dlatego trudno uznać je na obecnym etapie badań za wiarygodne biologicznie hipotezy. W niniejszej rozprawie przyjęto, odnosząc powyższe rozważania do modelu w wersji 1.2., że za powstanie zachowań wyższego poziomu (Z) odpowiada podsystem zarządzania zachowaniami wyższego poziomu (P-ZZWP), który na podstawie zgromadzonych sekwencji wygenerowanych przez podsystem P-H-RL składających się z: reprezentacji stanów świata (s_t, s_{t+1}), zachowań elementarnych (z_t), błędu predykcji nagrody (δ_{t+1}) oraz korelacji między wymienionymi elementami ($s_t , z_t \rightarrow s_{t+1}, \delta_{t+1}$) jest w stanie wytworzyć reprezentacje Z, tzw. opcje. Należy przyjąć, zgodnie z sugestiami specjalistów zajmujących się uczeniem maszynowym, że podsystem P-ZZWP, na podstawie napływającego strumienia danych, „próbuje" najpierw wyznaczyć abstrakcyjny model środowiska, w którym realizowane są działania, a następnie wyznacza na jego podstawie stany kluczowe dla tego modelu. W tak określonym układzie opcję rozumieć można jako

zachowanie, które pozwala agentowi przemieszczać się między wyróżnionymi stanami środowiska, np. wejściem do jakiegoś pomieszczenia oraz wyjściem (Lakshminarayanan i in., 2016). Agent może w istotny sposób ograniczyć proces eksploracji, dysponując tego typu opcjami, wystarczy, że rozpozna on w nim charakterystyczną strukturę (model), by móc skutecznie zastosować zachowania wyższego poziomu, które bez błędów przeprowadzą go między wyróżnionymi stanami środowiska.

Zachowania wyższego poziomu jako nagradzające akty behawioralne (R_Z) – (H3)

Wiedza dotycząca funkcjonowania nagród w układzie nerwowym jest rezultatem badań prowadzonych przez kilkadziesiąt lat (Berridge & Kringelbach, 2015). Z każdym rokiem coraz lepiej rozumiemy mechanizmy rządzące układem nagrody, ciągle jednak wiele istotnych problemów czeka na rozstrzygnięcie. Dotychczas udało się wyodrębnić i doprecyzować trzy główne aspekty funkcjonowania nagród: aspekt przyjemnościowy, aspekt motywacyjny i aspekt związany z oddziaływaniem nagród na uczenie się. Poczynione ustalenia pozwalają identyfikować mechanizmy wpływające na szczególną pozycję nagród w układzie nerwowym. Nadal jednak nie wiadomo, w jakich warunkach i w jaki sposób powstają nowe typy nagród, w szczególności te, które wytwarzane są niezależnie od wrodzonej reakcji organizmu na dany obiekt, akt behawioralny czy stan wewnętrzny. Wiadomo, że zjawisko „uznawania" określonego obiektu, zachowania czy stanu wewnętrznego za nagrodę zależy w dużym stopniu od uczenia poprzez tworzenie asocjacji i związane jest z doznawaniem szeroko pojętej przyjemności (patrz: zachowania konsumacyjne takie jak: jedzenie, picie, akt seksualny). Podczas badań eksperymentalnych dotyczących sposobów reprezentowania odmiennych typów nagród, takich jak przekąska oraz kwota pieniędzy porównywalna z wartością przekąski, zauważono, że w obu przypadkach aktywny był ten sam podobszar brzuszno-przyśrodkowej kory przedczołowej (Levy & Glimcher, 2012). Podobne obserwacje przeprowadzono podczas eksperymentów na makakach, które były skłonne zrezygnować z cennej dla nich nagrody w postaci soku, jeśli zamiast tego mogły uczestniczyć w spotkaniu z osobnikiem stojącym wyżej w hierarchii. W przypadku, gdy spotkanie dotyczyło kogoś o niższym statusie, wówczas udział w nim musiał być poprzedzony „przekupstwem" w postaci dodatkowej porcji soku. Najwyraźniej tylko wybrane formy relacji społecznych były dla nich cenniejsze niż smakołyki. Ten i podobne eksperymenty pokazują, że zwierzęta – podobnie jak ludzie –

dysponują mechanizmami pozwalającym szacować różne typy nagród w taki sam sposób (Deaner i in., 2005). Na tej podstawie sformułowano hipotezę wspólnej neuronalnej waluty (*a neural common currency*) (Levy & Glimcher, 2012)), głoszącą, że wartość wszystkich typów nagród szacowana jest przez ten sam mechanizm, co prowadzi do ich uniwersalnej „wyceny". Zdaniem teoretyków ekonomii jest to sytuacja bardzo pożądana z perspektywy agenta, gwarantuje mu bowiem spójny system preferencji, a co za tym idzie – możliwość dokonywania racjonalnych wyborów traktowanych jako przedmiot badania w takich teoriach decyzji jak: teoria oczekiwanej użyteczności (Von Neumann, 1944), teoria perspektywy (Kahneman & Tversky, 1979) czy teoria uczenia się ze wzmacnianiem (Sutton, 1998)[124].

Gdy uznamy hipotezę „wspólnej neuronalnej waluty" za wiarygodną, to problem tworzenia nowych typów nagród będzie można sprowadzić do następującej reguły: dany obiekt, stan wewnętrzny, akt behawioralny uzyskują status nagrody wówczas, gdy odpowiadająca im reprezentacja uzyska w podsystemie sensoryczno-ewaluacyjnym niezerową wartość. Kiedy reprezentacja uzyska tego typu status, to w zasadniczy sposób zmieni się jej funkcja w układzie nerwowym. Można powiedzieć, stosując terminologię teorii intencjonalności, że tego typu reprezentacja zmienia swoje nakierowanie na zgodność (patrz: punkt *Struktura intencjonalności* zaprezentowana w rozdziale drugim). Początkowo posiada ona nakierowanie typu umysł→świat (np. udostępnia agentowi jakiś obiekt z otoczenia), by z czasem, gdy obiekt zacznie być traktowany jak nagroda, uzyskać nakierowanie typu: świat → umysł. Od tego momentu aktywowanie tego typu reprezentacji będzie powodowało, że agent podejmie określone działania, których celem będzie dostosowanie świata do zaprojektowanego w umyśle i pożądanego przez agenta stanu, czyli do pozyskania określonej nagrody lub takiej zmiany otoczenia, by uniknąć negatywnego wpływu otoczenia, np. oddalić się od miejsca, w którym jest niebezpiecznie.

Obecnie nie wiemy, w jaki sposób funkcjonuje mechanizm odpowiedzialny za „wycenianie" reprezentacji i przekształcanie ich w nagrody. Niektórzy badacze sugerują, że jest on związany z energetycznymi potrzebami agenta (Montague, 2006). Wszystko, co w jakiś sposób powoduje nagły spadek zasobów energetycznych (np. trucizna, stres) lub umożliwia ich uzupełnienie (pożywienie) – z pewnością nie jest obojętne dla organizmu

[124] Tezę o tym, że wartość wszystkich typów nagród szacowana jest przez ten sam mechanizm, co prowadzi do ich uniwersalnej wyceny sformułował Paul Samuelson w pracy opublikowanej w 1947 roku, zatytułowanej *Foundations of Economic Analysis* (za Levy & Glimcher, 2012).

(Montague, 2006). Trudno jednak, na podstawie aktualnej wiedzy, powiedzieć, jak z podstawowego poziomu funkcjonowania organizmu przejść do nagród związanych z prestiżem, chęcią zdobycia sławy, itp. Jedno wydaje się pewne, by repertuar zachowań celowych danego agenta mógł się zwiększać, organizm powinien dysponować mechanizmem kreującym nowe typy nagród. Z perspektywy przetrwania, możliwość dostosowania organizmu do zmieniających się warunków otoczenia przez kreowanie nowych nagród to ważne osiągnięcie ewolucyjne, zasadniczo zwiększające zdolności agenta i pozwalające mu wykroczyć poza repertuar zachowań wrodzonych. Natomiast z perspektywy ontogenetycznej wrodzone oraz nabyte formy zachowań tworzą architekturę samowsporną (*bootstrapping*). Początkowo, większość potrzeb zaspokajana jest przy wykorzystaniu zachowań wrodzonych, motywowanych potrzebami (nagrodami) odziedziczonymi po przodkach w ramach ewolucji danego gatunku. Zapewniają one agentowi, że nie musi on od początku wypracowywać własnego systemu nagród, własnych preferencji dotyczących tych elementów środowiska, które są mu niezbędne do przeżycia, gdyż te, które odziedziczył, w dużym stopniu są „skalibrowane" i gotowe do wykorzystania. Z czasem, w trakcie rozwoju osobniczego, zachowania i preferencje agenta zostają poszerzone o nowe typy nagród i mogą obejmować:

i. szeroką klasę pokarmów dostępnych w środowisku z uwzględnieniem specyficznych dla gatunku ograniczeń (patrz: podział na roślinożerców, mięsożerców i wszystkożerców),
ii. zbiór stanów wewnętrznych (np. doznanie przyjemnego ciepła podczas kąpieli w gejzerze, emocję radości wywołaną zabawą),
iii. różnorodne akty behawioralne (np. oszustwo taktyczne (*tactical deception*), różne formy odstraszania przeciwnika czy przyciągania uwagi partnera lub partnerki).

Charakterystyczna dla danego gatunku różnorodność zachowań realizowanych ze względu na rozmaite nagrody jest pochodną dwóch procesów: (1) procesu eksploracji oraz (2) procesu wyceniającego reprezentacje. Im większa skłonność danego gatunku do testowania różnych stanów środowiska, im szersza gama pozytywnych lub negatywnych wrażeń wywołanych przez oceny napotkanych obiektów, przeżytych doświadczeń oraz zrealizowanych lub zaobserwowanych aktów behawioralnych, tym pojemniejsza baza nagród, a w konsekwencji – większa różnorodność celów i zachowań. Warto przypomnieć,

że mechanizm uczenia się ze wzmacnianiem zawiera w sobie dwa tryby działania: (1) eksplorację, wykorzystującą m.in. losową strategię wyboru zachowań (stosowaną głównie w początkowych fazach uczenia się) oraz (2) eksploatację, która losowy wybór działań zastępuje wyborami opartymi na funkcji wartości (skonstruowanej na podstawie wiedzy pozyskanej w fazie eksploracji środowiska). W systemach sztucznych „moment" przełączenia z pierwszego trybu w drugi przebiega zgodnie z pewną funkcją, która stopniowo zmniejsza udział czynnika losowego w wyborze działań. Można powiedzieć, odnosząc wskazane tryby do problemu różnorodności zachowań oraz typów nagród nabywanych przez agenta w toku rozwoju ontogenetycznego, że gatunki, które stosują krótką fazę eksploracji i szybko przechodzą do fazy eksploatacji (tzw. strategii zachłannej), mają ograniczony zbiór nagród, które w przyszłości będą wpływały na ich zachowania. Proces wyceny, która jest „wrażliwa" jedynie na wąski zakres reprezentacji, wywołuje podobny efekt. Innymi słowy, jeśli tego typu mechanizm nie będzie reagował na szeroki zakres doświadczeń (smaków, zapachów, wrażeń, itp.), to trudno oczekiwać, by agent wykroczył poza zbiór typowych dla niego nagród oraz związanych z nimi zachowań.

Przedstawiony powyżej mechanizm konstruowania nowych typów nagród tłumaczy zachowania nabyte, które powiązane są z przetrwaniem w środowisku. Pojawia się jednak pytanie: czy jest on wystarczający, aby można było wykorzystać go do bardziej abstrakcyjnych przypadków, takich choćby jak akty behawioralne oparte na wyróżnionych wcześniej zachowaniach wyższego poziomu (Z)? Należy przypomnieć, że akty behawioralne są jednym z wielu typów nagród. Problem polega na tym, że zachowania wyższego rzędu to reprezentacje abstrakcyjne, które nie zawsze dają się w prosty sposób wycenić za pomocą standardowych mechanizmów. Wydaje się, że w takim przypadku potrzebny jest mechanizm w obrębie modułu ewaluacyjnego, który „nauczy się" wyceniać tego typu reprezentacje.

Z podobną sytuacją mamy do czynienia w przypadku procesu uczenia się funkcji wartości wykorzystywanej przez algorytm TDRL. Początkowo wartość tej funkcji jest dla danego stanu zerowa lub losowa, dopiero w trakcie kolejnych interakcji ze środowiskiem funkcja ta – poprzez stopniowe korekty realizowane przy uwzględnieniu błędu predykcji nagrody (δ) – zaczyna prawidłowo reprezentować zdyskontowaną, oczekiwaną sumę przyszłych nagród, niezbędną do wyznaczenia optymalnej strategii zachowań. Można założyć, że podobnie przebiega proces szacowania wartości zachowań wyższego poziomu. Początkowo może być ona losowa lub arbitralnie wyznaczona na podstawie pierwszego

doświadczenia. Z czasem jednak wartość zachowania Z, po wykonaniu szeregu powtórzeń wyceny uzyskanych nagród oraz zrealizowaniu odpowiednich korekt, zacznie zbliżać się do wartości oczekiwanej, uwzględniającej różne konteksty oraz możliwości jej pozyskania. Na przykład, zachowanie polegające na przyciąganiu uwagi przyszłej partnerki może mieć bardzo wysoką wartość, kiedy realizowane jest w okresie godowym, a równocześnie być nisko wyceniane, kiedy odbywa się w czasie, gdy osobnik płci przeciwnej nie jest zainteresowany zalotami.

Z dotychczasowych rozważań wyłania się następujący obraz. Zachowania wysokiego poziomu realizują dwie ważne funkcje: (1) ograniczają koszty eksploracji poprzez uogólnienie nabytych doświadczeń, co w konsekwencji prowadzi do bardziej optymalnej realizacji celów oraz (2) uzyskują status nagradzających aktów behawioralnych, co znaczy, że realizacja samych tych aktów odbierana jest przez agenta jako coś wartościowego, co w określonych okolicznościach należy wykonać. Aby powyższe funkcje mogły być zrealizowane, wymagane są również odpowiednie rozszerzenia w pozostałych podsystemach. P-SEUR musi nauczyć się wyznaczać R_Z dla danego Z. Z kolei podsystem monitorowania i motywacji powinien wiązać nagrody R_Z z obserwacjami (o/O), które towarzyszą realizacji działania Z. Wreszcie P-H-RL powinien być w stanie inicjować realizację zachowania na wyższym poziomie w kontekście obu wymienionych funkcji, tzn. jako cel sam w sobie oraz jako element wspomagający realizację innych celów. Powiedzieć można, że pojawienie się podsystemu zarządzania zachowaniami wyższego poziomu wpływa na modyfikację funkcjonowania pozostałych podsystemów.

Przedstawione powyżej mechanizmy tworzenia, stosowania i ewaluacji zachowań wysokiego poziomu Z są na obecnym etapie badań „mieszanką" neuronaukowych danych empirycznych, obserwacji behawioralnych i spekulacji teoretycznych. Ich wartość polega głównie na tym, że pokazują one, z jakimi problemami musi zmierzyć się badacz „obliczeniowo zorientowanej" kognitywistyki. Jest to szczególnie istotne, gdy zaczyna on/ona rozważać reprezentacje odnoszące się nie tyle do pojedynczych doświadczeń, co do ich uogólnionej postaci. Abstrakcyjny charakter zachowań wyższego rzędu (Z), kontekstowa niezmienność obserwacji (O), uśredniona wartość nagrody aktu behawioralnego (R_Z) to ważne narzędzia opanowywania i radzenia sobie z często nieprzewidywalnymi zmianami otoczenia. Scharakteryzowane wyżej mechanizmy umożliwiają rozpoznawanie różnego rodzaju wzorców, prawidłowości i reguł postępowania, które włączone odpowiednio w podsystem kontroli zachowań pozwalają

skuteczniej zaspokajać potrzeby agenta. Nowe typy reprezentacji nie tylko pozwalają lepiej uchwycić określone związki istniejące w świecie, ale również pozwalają one uwolnić się agentowi od podporządkowania temu, co dzieje się tu i teraz, czyli sprawiają, że ma on do dyspozycji informacje wykraczające poza raporty o bieżącym stanie środowiska[125].

Jeśli organizm nie dysponuje kluczowymi dla podjęcia danej decyzji informacjami, to albo jest on skazany na popełnianie błędów, których nigdy nie wyeliminuje, albo uda mu się w jakiś sposób uzupełnić niedoskonałości bezpośredniej obserwacji. Warto rozważyć następujący hipotetyczny przykład[126], aby podkreślić znaczenie tego dylematu. Wyobraźmy sobie, że pewien gatunek zwierząt nauczył się rozpoznawać ślady zostawiane przez polujących na niego drapieżników. Tego typu umiejętność w istotny sposób zmniejsza ryzyko śmierci przedstawicieli tego gatunku, z drugiej strony – zbyt asekuracyjne reagowanie na ślad drapieżnika może pozbawić jego osobniki szans na pozyskanie pożywienia z tych terenów, po których przemieszcza się drapieżnik. Optymalne rozwiązanie polegałoby nie tylko na rozpoznaniu śladu drapieżnika, ale również na określeniu czasu jego powstania. Jeśli ślad jest świeży, to lepiej zrezygnować z dalszej eksploracji danego obszaru. Jeśli jednak ślad jest stary, to ryzyko zagrożenia jest niewielkie, a więc można je zignorować. Wskazana sytuacja wymaga od zwierzęcia wyjścia poza aktualne informacje o jego otoczeniu. By podjąć prawidłową decyzję – zostać czy opuścić dany teren, trzeba odwołać się do pewnej uogólnionej reprezentacji śladów drapieżnika i dodatkowo powiązać je z bieżącym stanem środowiska. Tego typu zdolność, jak pokaże model 2.0, jest jedną z podstawowych cech ludzkiego umysłu. Człowiek, już na wczesnym etapie swojego rozwoju, posługuje się reprezentacjami uogólniającymi wcześniejsze doświadczenia i przeprowadza wnioskowania, które rekompensują braki bezpośredniej obserwacji.

Przedstawione wyżej modele od 1.0 do 1.2 charakteryzowały działanie intencjonalne tylko ze względu na Cechę 1[127]. Kolejne konkretyzacje modelu 1.0 wprowadzały czynniki

[125]Algorytm uczenia się ze wzmacnianiem zakłada, że reprezentacja stanów środowiska opiera się na tzw. własności Markowa. Tylko przy tym założeniu podstawowy algorytm RL gwarantuje zbieżność, czyli możliwość osiągnięcia optymalnej polityki doboru zachowań. Bywa jednak tak, że wskazana własność jest trudna do osiągnięcia i w związku z tym stosuje się tzw. aproksymatory funkcji wartości.

[126] Przytoczony przykład inspirowany jest badaniami nad koczkodanami zielonosiwymi (Chlorocebus aethiops), które wykazały, że małpy tego gatunku potrafią rozróżniać i komunikować pozostałym członkom stada wykryty gatunek drapieżnika (lampart, wąż, orzeł) (Seyfarth i in., 1980).

[127] Cecha 1: Zależność od kontekstu i wyuczonych wcześniej asocjacji. „Działania intencjonalne tylko w niewielkim stopniu zależą od bezpośrednich bodźców, zaś w dużym stopniu zależą od kontekstu zadania oraz od wyuczonych wcześniej powiązań.” (Haggard, 2005, s. 291).

będące dookreśleniem tej cechy. Jednak żaden z powyższych modeli nie uwzględniał tego, że konstrukcja oraz realizacja działań intencjonalnych wymaga od agenta skupienia uwagi oraz aktywnego monitorowania czy uzyskane efekty zgodne są z zamiarami. Dalsze urealistycznienie modelu działania intencjonalnego wymaga zatem włączenia do niego składników, które spełniałyby warunki sformułowane w Cesze 2. Kolejny model uwzględniać zatem powinien rolę sieci procesów poznawczych w planowaniu i kontroli działania intencjonalnego. Na silną zależność między działaniami intencjonalnymi a procesami poznawczymi zwraca również uwagę Patrick Haggard (Haggard 2005, s. 291). Polega to na tym, że w planowanie i skuteczną realizację działania zaangażowany jest cały szereg elementarnych procesów poznawczych (głównie uwaga, percepcja, pamięć, czyli tzw. funkcje wykonawcze), a na odpowiednim etapie rozwoju osobniczego dołączają do nich złożone procesy poznawcze, takie jak: rozumowanie, jawne wartościowanie, przewidywanie, symulowanie w umyśle możliwych skutków działania, itp. Oczywiste jest, że te zaawansowane procesy poznawcze wymagają dysponowania odpowiednio rozwiniętą kompetencją językową (Nęcka i in., 2006b). Niestety, wiedza na temat funkcjonowania tego typu procesów nie została do tej pory efektywnie uwzględniona w pracach z psychologii intencji (patrz: rozdział czwarty) ani w badaniach dotyczących neurobiologicznych podstaw procesów decyzyjnych (patrz: rozdział trzeci). Natomiast filozoficzna teoria intencjonalności Searle'a, choć zakłada się w niej istnienie ścisłego związku między zachowaniem a zamiarem, czyli stanem należącym do sieci stanów intencjonalnych (patrz: podstawowy schemat przebiegu działania intencjonalnego zaprezentowany w rozdziale 2.), to jednak nie precyzuje się w niej, które procesy poznawcze i w jakim zakresie są niezbędne do skonstruowania takiego zamiaru i jego zrealizowania. Wypełnienie tak istotnej luki to zadanie wymagające zaangażowania całych grup badaczy. Dlatego też w proponowanych w niniejszej pracy modelach działań intencjonalnych nie uwzględnia się konkretnych procesów poznawczych związanych z realizacją takich działań, lecz zakłada się jedynie globalny efekt funkcjonowania tego typu procesów. Efekt ten wykorzystywany jest w tworzeniu planu realizacji działania intencjonalnego. Uważam, iż pomimo tego ograniczenia, jakim jest pominięcie udziału konkretnych procesów poznawczych w działaniu intencjonalnym, nadal możliwa jest odpowiedź na pytanie: **jaką strukturę musi posiadać system kontroli zachowań, skoro wbudowane są w niego dwa - do pewnego stopnia konkurujące ze sobą - mechanizmy selekcji działań**: pierwszy, oparty na metodzie prób i błędów (a więc bez udziału procesów poznawczych) oraz drugi, oparty na planowaniu nadzorowanym przez procesy poznawcze?

Odpowiedź na tak postawione pytanie będzie rezultatem dociekań w kolejnych podrozdziałach. W trakcie omawiania modelu 2.0 wykorzystane zostaną nie tylko analizy Searle'a oraz zasada działania mechanizmu uczenia się ze wzmacnianiem, ale również odpowiednio zinterpretowane wyniki badań przeprowadzonych przez psychologów intencji[128].

[128] Zgodnie z ustaleniami referowanymi w rozdziale czwartym do grona psychologów intencji zalicza się m.in.: Benjamina Libeta, Patricka Haggarda, Daniela Wegnera, Susan Pocket.

5.3.4 Model 2.0 – działanie intencjonalne osadzone w sieci procesów poznawczych

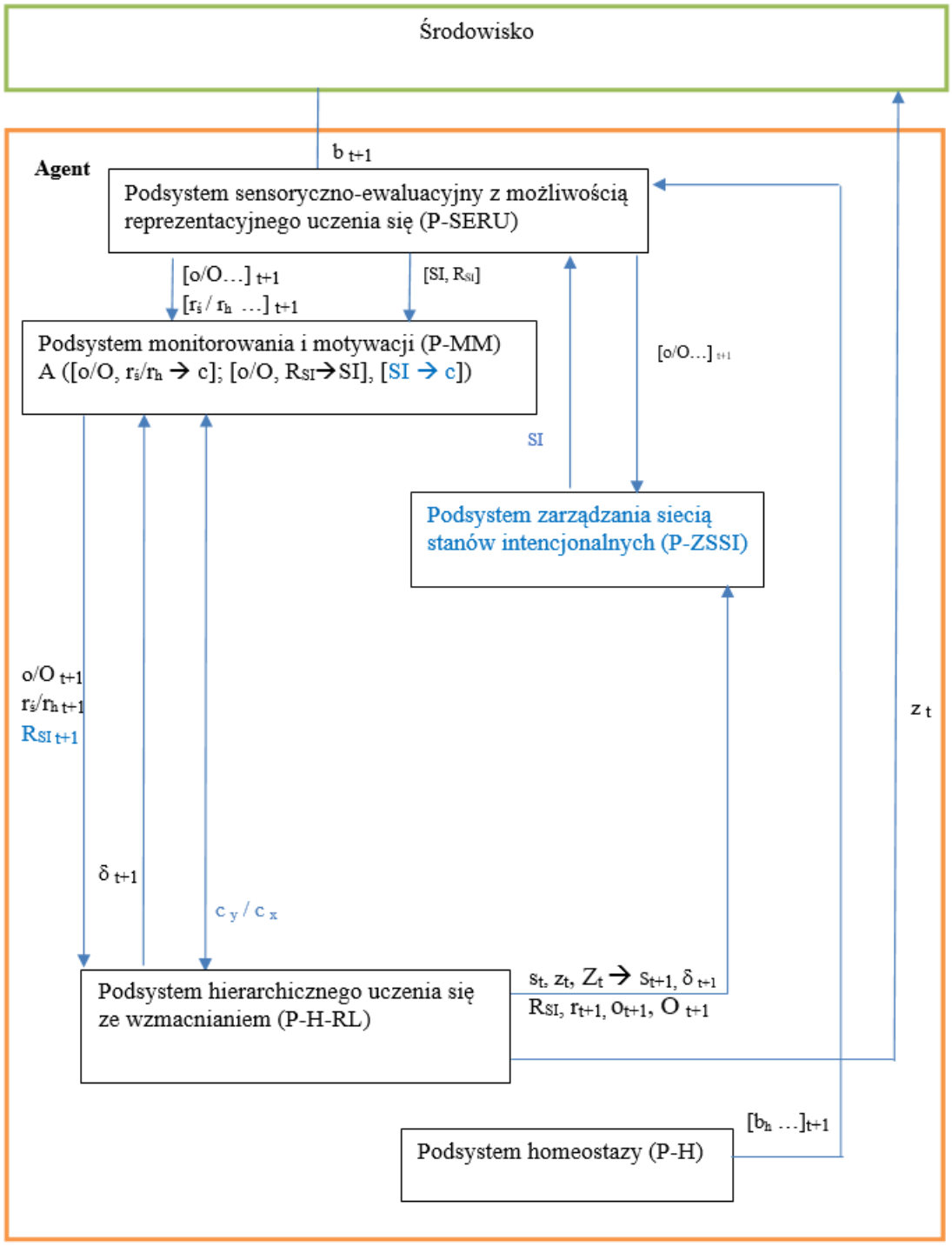


Diagram 13. Model działania intencjonalnego z podsystemem zarządzania siecią stanów intencjonalnych (P-ZSSI).

Legenda symboli (uzupełnienie w odniesieniu do Modelu 1.2):

- SI – reprezentacja stanu intencjonalnego utworzona w ramach podsystemu zarządzania siecią stanów intencjonalnych; stan intencjonalny w ujęciu Searle'a posiada strukturę, która decyduje o tym, w jaki sposób odnosi się on do rzeczywistości i jaką treść zawiera; równocześnie poszczególne stany wzajemnie się warunkują i tworzą ze sobą sieć relacji;
- R_{SI} – reprezentacja wartości nagrody związanej z danym stanem intencjonalnym SI; wartość R_{SI} wyznaczana jest – podobnie jak w poprzednich modelach – przez P-SERU, zgodnie z hipotezą wspólnej neuronalnej waluty.

Uzasadnienie

Przedstawiony powyżej diagram Modelu 2.0 zawiera jedno kluczowe rozszerzenie w porównaniu z wcześniejszymi modelami. Polega ono na zastąpieniu podsystemu zarządzania zachowaniami wyższego poziomu (P-ZZWP) przez podsystem zarządzania siecią stanów intencjonalnych (P-ZSSI). W przyjętej tu koncepcji P-ZSSI jest uogólnieniem i rozszerzeniem P-ZZWP. Główną funkcją wskazanego rozszerzenia jest wzbogacenie mechanizmu kontroli zachowań o możliwość wykorzystywania wiedzy dziedzinowej, którą zgromadzono w sieci stanów intencjonalnych.

Można powiedzieć, nawiązując do wprowadzonego przez Searle'a pojęcia tła, że wcześniejszy Model 1.2 dostarczał takie wyjaśnienie działania intencjonalnego, które uwzględniało różnego rodzaju dyspozycje tła (osadzenie działania w określonym kontekście to właśnie uwzględnienie wpływu, jaki wywiera tło), czyli było to wyjaśnienie zachowania celowego wykształconego metodą prób i błędów, realizowanego w związku z chęcią pozyskania przez agenta różnego rodzaju nagród (od zaspokojenia potrzeb biologicznych do potrzeb abstrakcyjnych). Tego typu dyspozycje pozwalają organizmowi skutecznie działać w środowisku, które jest w dużym zakresie do niego dostosowane (tzw. nisza). Stosowany w tym przypadku mechanizm kontroli zachowań bazuje na bieżących obserwacjach oraz nagrodach, czyli natychmiastowej, wartościującej informacji zwrotnej, która czerpana jest ze źródeł dostępnych tu i teraz[129]. W rzeczywistym świecie wiele zjawisk jest na tyle złożonych, że bieżąca obserwacja oraz informacja wartościująca nie wystarczają, by trafnie przewidzieć przebieg danego zjawiska, a co za tym idzie, by

[129]Formalnie tego typu wymaganie to, przywoływana wielokrotnie, własność Markowa (P. Cichosz, 2007, s. 728).

dostosować zachowanie do jego zmian. Taka zdolność przewidywania jest szczególnie potrzebna w przypadku zachowań społecznych, w których wiedza o wcześniejszych działaniach członka stada lub społeczności jest niezbędna do efektywnej kooperacji. Jeśli, przykładowo, ktoś sprawia wrażenie, że chce nam pomóc, ale wcześniej wielokrotnie nas oszukał i wykorzystał, to raczej nie zdecydujemy się skorzystać z jego oferty. Dobre wrażenie, uśmiech, miłe słowa będą na ogół niewystarczające, aby na ich podstawie podjąć decyzję o pożyczeniu komuś takiemu pieniędzy. Niestety, metoda uczenia się ze wzmacnianiem nie sprawdza się w tego typu kontekstach. Wbudowane w nią założenie dotyczące modelu środowiska, w którym działa agent (tzw. proces decyzyjny Markowa), mocno komplikuje wykorzystanie tego typu informacji do konstrukcji optymalnej strategii doboru zachowań. By móc „skompensować" wskazane ograniczenie, konieczny jest mechanizm, który pozwoli włączyć w proces decyzyjny informację wykraczającą poza dostępne agentowi w danym stanie świata obserwacje (o/O). Możemy, posiadając „model oszusta" lub oszustwa, nie tylko przygotować się na różne formy manipulacji, ale również skutecznie się im oprzeć. Wskazany przykład obrazuje, jak odpowiednio zorganizowana sieć reprezentacji włączona w system kontroli zachowań może w radykalny sposób poprawić efektywność działania agenta. Znaczy to, że podsystem zarządzania stanami intencjonalnymi powinien być odpowiednio zintegrowany z podsystemem „programującym" dobór zachowań elementarnych. Nie musi to prowadzić do tego, że pomiędzy stanem intencjonalnym a określonym ruchem ciała (lub sekwencją ruchów) zachodzi relacja bezpośredniego związku przyczynowego, niemniej należy oczekiwać wpływu tego typu stanów na podsystem odpowiedzialny za dobór zachowań elementarnych lub zachowań wyższego poziomu.

Obecnie pogłębiona zostanie wstępna charakterystyka nowych podsystemów, którą przedstawiłem powyżej. Skupię się przede wszystkim na udzieleniu odpowiedzi na dwa następujące pytania:

- W jaki sposób interakcje ze światem determinują kształt sieci stanów intencjonalnych?
- W jaki sposób wiedza zawarta w sieci stanów intencjonalnych wpływa na kontrolę zachowań?

W jaki sposób interakcje agenta ze światem determinują kształt sieci stanów intencjonalnych?

Podczas prezentacji modelu 1.2 zwrócono uwagę na to, iż posiadanie zdolności do wytwarzania nowych typów reprezentacji pozwala wypracować nowe sposoby oddziaływania na środowisko. Skuteczne identyfikowanie źródeł danych sensorycznych, niezależnie od kontekstu (obserwacje typu O) czy wykorzystywanie umiejętności opanowanych w jednej dziedzinie do realizowania zachowań wyższego poziomu w innej dziedzinie (Z) (np. posługiwanie się wyuczoną formą percepcji wzrokowej w procesie czytania) są przykładami tego, jak agent może zwiększać efektywność swojego działania poprzez użycie nowych typów reprezentacji i to zarówno w środowisku naturalnym, jak i społecznym (mam tu na uwadze głównie reprezentacje językowe, za pomocą których można wpływać na zachowania innych ludzi). Tego typu reprezentacje łączą element konkretnego doświadczenia z jego uogólnieniem. Konstrukcja wymienionych typów reprezentacji jest przejawem pewnej prawidłowości funkcjonowania układu nerwowego polegającej na „pomijaniu" nieistotnych dla danego zjawiska szczegółów i „skupianiu się" na najważniejszych i uniwersalnych jego cechach – na pewnym wzorcu. W eseju dotyczącym problemu świadomości Karl Friston zauważa:

> *Z każdym nowym doświadczeniem twój organizm przeprowadza wnioskowanie, aby dopasować to, czego w danej chwili doświadcza, do znanego wzorca.*[130]

Identyfikowanie ogólnych prawidłowości w napływających danych jest pierwszą funkcją podsystemu odpowiedzialnego za konstrukcję reprezentacji, przyjmujących formę złożonej sieci stanów intencjonalnych (P-ZSSI).

Wydobywanie informacji z napływających danych sensorycznych oraz „poszukiwanie" zależności istniejących między faktami jest drugim ważnym zadaniem procesów odpowiedzialnych za zarządzanie stanami intencjonalnymi. Warto w tym kontekście przypomnieć spostrzeżenie Freda Dretskego:

> *[…] zobaczyć więcej faktów [można] nie tylko dzięki postrzeganiu większej liczby przedmiotów, lecz dzięki rozszerzaniu wiedzy na temat tego, co postrzegane już przedmioty wyrażają na temat przedmiotów, których zobaczyć nie można*. (Dretske, 2004, s. 56).

[130] „With every new experience, your organism engages in inference to fit what's happening into a familiar pattern." (Friston, 2017).

Wskazana zasada, by w strumieniu[131] napływających reprezentacji identyfikować istniejące między nimi związki, odpowiedzialna jest nie tylko za tworzenie połączeń w sieci stanów intencjonalnych oraz ich gęstość, ale również za jej holistyczny charakter.

Zdolność do abstrahowania oraz zdolność do tworzenia złożonych sieci znaczeń – z perspektywy kontroli zachowań – to dwie najistotniejsze cechy stanów intencjonalnych[132]. Pozwalają one podmiotowi konstruować pojęciowe modele określonych domen rzeczywistości po to, aby posłużyć się nimi do jej zmiany. Zaprezentowana poniżej propozycja stopniowego konstruowania sieci stanów intencjonalnych stanowi wstępną syntezę teorii intencjonalności Searle'a, badań psychologów intencji (Daniela Wegnera i Patricka Haggarda) oraz wybranych elementów teorii kodowania predykcyjnego Karla Fristona.

Warto przypomnieć, że sieć stanów intencjonalnych jest wytworem dyspozycji tła oraz związanych z nimi niejawnych, przedintencjonalnych oczekiwań i postaw (*stance*) wobec rzeczywistości. Są to przejawy tzw. wiedzy-jak, obejmującej wiedzę o tym, jakie są rzeczy oraz jak coś z ich pomocą wykonać. Tego typu „wiedza", w opinii Searle'a, ma charakter niereprezentacyjny, równocześnie jednak warunkuje ona poziom intencjonalny, zabezpieczając go przed regresem w nieskończoność. Odnosząc relację tło-stany intencjonalne-język w pewnym miejscu Searle stwierdza:

> *Jeśli reprezentacja zakłada Tło, to samo to Tło nie może się składać z reprezentacji bez generowania nieskończonego regresu. Wiemy, że nieskończony regres jest empirycznie niemożliwy, ponieważ ludzkie możliwości intelektualne są ograniczone. Sekwencja kroków poznawczych niezbędnych do zrozumienia wypowiedzi językowej* [nadbudowanej nad siecią stanów intencjonalnych – uwaga M.C.], *musi w pewnym*

[131] W informatyce strumień jest sekwencyjną strukturą danych udostępniającą wchodzące w jej skład elementy na żądanie (Bewig, 2007).

[132] Warto w tym miejscu dodać, że na bazie tych samych doświadczeń implementowana jest również perspektywa egocentryczna, w kontekście której wiele świadomych przeżyć zyskuje status przeżyć fenomenalnych o unikatowym , jakościowym charakterze (Nagel, 2012). W omawianych zjawiskach szczególnie widoczne są indywidualne cechy poszczególnych doświadczeń, ich niepowtarzalność i swoistość. Sytuacja komplikuje się jeszcze bardziej, zdaniem Searle'a, kiedy dodatkowo uwzględni się różne czynniki warunkujące nasze przeżywanie świata. Mowa o nastrojach, zdolności do strukturyzowania danych zmysłowych (problem tła i figury), polu uwagi z jej topologią opartą na centrum i peryferiach, o usytuowaniu reprezentacji w szerszym kontekście przestrzenno-czasowym (*situatedness*), o aspektowości stanów intencjonalnych (*aspectual shape*), o ich emocjonalnym nacechowaniu (bycie podekscytowanym lub znudzonym), o związkach z językiem czy wreszcie o jedności umysłu, która łączy w całość wszystkie wymienione wymiary strumienia przeżyć – i to zarówno horyzontalnie (organizacja doświadczenia w wymiarze czasowym), jak i wertykalnie (doświadczane symultanicznie w danej chwili wielorakie odniesienie przedmiotowe, np. dźwięki, obiekty znajdujące się w polu percepcyjnym) (Searle, 1992, s. 130).

> *momencie dobiec końca. Zgodnie z przedstawioną tu koncepcją, to [językowe rozumienie] nie kończy się wraz z uchwyceniem wyizolowanej treści semantycznej, a nawet treści semantycznej wraz z towarzyszącym jej zbiorem uprzednich przekonań. Jest raczej tak, że treść semantyczna funkcjonuje jedynie w kontekście Tła, które składa się z kulturowego i biologicznego know-how i to właśnie to know-how Tła umożliwia nam zrozumienie dosłownych znaczeń.*[133]

Trudno się zgodzić z tezą Searle'a, że dyspozycje tła są całkowicie niereprezentacyjne. Jeśli twierdzi się, tak jak on, że należąca do tła „wiedza-jak" ma charakter niereprezentacyjny, to należałoby pokazać, jak z takiej niereprezentacyjnej wiedzy tła wyłania się „wiedza-że", oparta na stanach intencjonalnych. Searle nie dostarcza wyjaśnienia tego przejścia od jednego do drugiego typu wiedzy. Wskazana niejasność ma prawdopodobnie swoje źródło w przekonaniu Searle'a, że podstawowy mechanizm funkcjonowania tła polega na uczeniu się metodą prób i błędów, niewymagającą, jego zdaniem, żadnych reprezentacji w trakcie jej stosowania. Z perspektywy współczesnych badań nad metodą uczenia się ze wzmacnianiem wiemy, że nie jest to trafny pogląd. Nie ulega wątpliwości, że zgodnie z zaprezentowanymi wcześniej podstawami teoretycznymi uczenia się ze wzmacnianiem, stosowanie metody prób i błędów bez odwołania się do odpowiednich obserwacji, reprezentacji nagród, funkcji wartości czy reprezentacji stanów świata nie pozwoliłoby agentowi na realizację wybranych przez niego celów. Można stwierdzić, że stosowane w informatyce podejście do uczenia się ze wzmacnianiem nie tylko pozwala rozwiązywać złożone obliczeniowo problemy, ale pomaga również w zrozumieniu, sugerowanej przez Searle'a, relacji pomiędzy Tłem a siecią stanów intencjonalnych, tj. pomiędzy rzekomo niereprezentacyjnym know-how dotyczącym świata a reprezentacyjną wiedzą zgromadzoną w sieci stanów intencjonalnych. Zgodnie bowiem z podejściem obliczeniowym zarówno na najniższym poziomie (wiedza-jak ukształtowana w dużym stopniu przy pomocy metody uczenia się ze wzmacnianiem na bazie reprezentacji wrodzonych), jak i na najwyższym (wiedza-że zorganizowana w formie sieci stanów intencjonalnych) mamy do czynienia z hierarchią reprezentacji o różnym

[133] „If representation presupposes a Background, then the Background cannot itself consist in representations without generating an infinite regress. We know that the infinite regress is empirically impossible because human intellectual capacities are finite. The sequence of cognitive steps in linguistic understanding comes to an end. On the conception presented here, it does not come to an end with the grasp of semantic content in isolation or even with semantic content together with a set of presupposed beliefs, but rather the semantic content only functions against a Background that consists of cultural and biological know-how, and it is this Background know-how which enables us to understand literal meanings" (Searle, 1983, s. 148).

stopniu złożoności. Aby rozpoznać cechy wskazanej hierarchii, należy porównać ze sobą następujące typy reprezentacji:

1. reprezentację obserwacji wykorzystywanej przez algorytm uczenia się ze wzmacnianiem (‘o’);
2. reprezentację obserwacji niezależnej od kontekstu, której model obliczeniowy opiera się na kodowaniu predykcyjnym (‘O’);
3. percepcyjny stan intencjonalny (SI_P) zakładany w teorii intencjonalności.

Każda z wymienionych reprezentacji odnosi się w podobny sposób do odpowiadającego jej stanu świata, tzn. posiada podobnie zorganizowany zbiór warunków spełniania (*condition of satisfaction*). – Warto przypomnieć, że w Searle’owskiej teorii intencjonalności o tego typu reprezentacjach mówi się, że posiadają nakierowanie na zgodność typu umysł→świat, tzn., że to na nich „spoczywa odpowiedzialność”, by wiernie odnosić się do świata. Główna różnica między nimi polega na zakresie udostępnianej przez reprezentację informacji, a w związku z tym – na ich uniwersalności i użyteczności.

W przypadku obserwacji typu ‘o’ mamy do czynienia z prostym pobudzeniem układu sensorycznego, które pozwala agentowi określić, w jakim stanie świata (s_t) się znajduje i na tej m.in. podstawie wybrać najbardziej korzystne z tej perspektywy działanie. Odniesienie przedmiotowe tego typu reprezentacji jest wąskie i w zasadzie sprowadza się do detekcji ściśle określonych stanów (cech) środowiska. Przykładem tego typu reprezentacji mogą być:

- informacja o wykryciu pewnej substancji w otoczeniu, o ile poziom jej stężenia przekroczy X jednostek
- informacja o dźwięku o częstotliwości Y,
- informacja o temperaturze obiektu znajdującego się w pobliżu, o ile jest ona większa od wartości Z.

Ich główną funkcją jest „zasilanie” algorytmu TDRL informacjami o stanie środowiska. Istotnym ograniczeniem tego typu reprezentacji jest ich niska zawartość informacyjna. W rezultacie, stworzony na ich podstawie model środowiska jest odpowiednio uproszczony, „niewrażliwy” na szereg dostępnych dla organizmu możliwości, np. alternatywnych typów pożywienia, szans na uniknięcie niebezpieczeństw poprzez wykorzystanie struktury terenu, itp. Tego typu model cechuje niska zdolność do różnicowania reprezentowanych przez niego zjawisk – różne zdarzenia/obiekty generujące podobne sygnały będą traktowane w

taki sam sposób. Łatwo wyobrazić sobie, że tego typu dwuznaczne przypadki będę prowadziły do zachowań nieefektywnych, a czasami nawet zagrażających życiu agenta, np. brak możliwości rozpoznania roślin owadożernych dla wielu owadów kończyłby się śmiercią.

Organizmy wyposażone w układ nerwowy dysponujący bardziej złożonymi formami przetwarzania i reprezentowania dostępnej w środowisku informacji są w stanie przezwyciężyć wskazane problemy, tworząc reprezentacje typu ‘O’. Za powstanie tego typu reprezentacji odpowiadają, zgodnie z koncepcją Fristona, dwa współpracujące ze sobą modele: model rozpoznawczy wykorzystujący informacje sensoryczne (te same, które są podstawą do konstrukcji reprezentacji typu ‘o’) oraz model generatywny oparty na wewnętrznych stanach organizmu, za pomocą którego dokonywane są predykcje dotyczące pobudzeń sensorycznych. Łączący wymienione modele mechanizm minimalizacji błędu predykcji powoduje, że poszczególne reprezentacje stają się z czasem bardziej uniwersalne, niezależne od kontekstu, „odporne” na szum zawarty w dochodzących do układu nerwowego sygnałach. Z perspektywy mechanizmu uczenia się ze wzmacnianiem dysponowanie tego typu reprezentacjami jest ważne. Reprezentacje stanów świata, które są tworzone na ich podstawie, są w konsekwencji stabilniejsze i uniwersalne, a przez to znacznie bardziej użyteczne, gdyż umożliwiają agentowi utworzenie adekwatnego i bardziej precyzyjnego modelu środowiska, nawet gdy dostępne informacje są szczątkowe lub do pewnego stopnia zniekształcone. Decyzje podejmowane na podstawie tego typu modelu będą efektywniejsze, a liczba popełnianych błędów mniejsza.

Wskazane cechy reprezentacji typu ‘O’ nie wyczerpują wszystkich możliwości reprezentowania świata, w szczególności przez mózg-umysł człowieka. Percepcyjne stany intencjonalne to niewątpliwie reprezentacje o najbardziej złożonej strukturze, które niosą informacje o intencjonalnych obiektach, a nie o cechach bodźców sensorycznych. Można zidentyfikować kilka istotnych rozszerzeń, porównując cechy percepcyjnego stanu intencjonalnego z odpowiadającą mu reprezentacją typu ‘O’. Pierwsze rozszerzenie dotyczy warunków spełniania, które odpowiadają za samoodniesienie przyczynowe. W ujęciu Searle’a stan percepcyjny odnoszący się do jakiegoś obiektu nie tylko udostępnia sam ten obiekt, ale również zawiera informację o tym, że ten obiekt był przyczyną powstania określonego przeżycia percepcyjnego. Reprezentacja typu ‘O’ nie zawiera w sobie tego typu informacji, w jej treści pojawia się jedynie obiekt, dla którego zachodzi zgodność danych sensorycznych oraz wygenerowanych przez model generatywny

predykcji. Ta, z pozoru niewielka, różnica pozwala agentowi odróżnić przedmioty dostrzeżone od wyobrażonych. Jest to zatem ważny element różnicujący stany intencjonalne w obrębie sieci, w której funkcjonują, a równocześnie znak, że agent dysponuje dodatkowymi zasobami poznawczymi (zaawansowaną pamięcią epizodyczną, wyobraźnią, zdolnością do myślenia kontrfaktycznego).

Forma aspektowa stanów intencjonalnych jest kolejnym rozszerzeniem tego typu reprezentacji w porównaniu z reprezentacjami typu ‘O’. Każdy stan intencjonalny, przywołując myśl Searle’a, udostępnia swój przedmiot tylko pod pewnym względem, np. kołyszących się rytmicznie ludzi postrzegamy w zależności od kontekstu jako tancerzy lub jako osoby będące pod wpływem alkoholu, lub jako sportowców rozgrzewających się przed biegiem. Już Frege twierdził, że jedno znaczenie (odniesienie przedmiotowe) może być ukazane za pomocą różnych sensów, a każdy z nich porównać można do punktu widzenia, który odsłania inny aspekt tego samego obiektu (Frege, 1977, s. 60). Często w tym kontekście mówi się o swoistej nieprzejrzystości treści stanów intencjonalnych, które wyodrębniają dany przedmiot jedynie pod pewnym względem (Gut, 2015). Tego typu efekt, jak twierdzi Searle, jest ściśle powiązany ze zdolnością podmiotu do kategoryzacji. Aby podmiot poznający mógł potraktować coś jako samochód, wcześniej powinien dysponować odpowiednią kategorią tego typu obiektów. Ponadto, jak twierdzi Searle, wybór kategorii jest w jakimś stopniu kontrolowany przez agenta. Kiedy patrzę na jakiś obiekt, to mogę spojrzeć na niego w różny sposób, mogę zaklasyfikować go do różnych kategorii. W ten sposób stany intencjonalne niejako dostosowują się do bieżących potrzeb agenta, eliminując (filtrując) nadmiar informacji zbędnych w danym kontekście środowiskowym.

Czysto przyczynowe teorie intencjonalności (np. funkcjonalizm) pomijają, zdaniem Searle’a, aspektowość stanów intencjonalnych.

> *Problem polega na tym, że przyczynowa teoria intencjonalności, jak choćby ta, którą rozwijają funkcjonaliści, nie obejmuje różnic w postaciach wyglądowych [aspektowych] konkretnych stanów intencjonalnych, ponieważ w związkach przyczynowo-skutkowych takich różnic nie ma. Czegokolwiek przyczyną jest woda, tego też jest przyczyną H_2O, a wszelkie związki przyczynowo-skutkowe, których skutkiem jest woda, obejmują w ten sam sposób H_2O* (Searle, 2010b, s. 98).

Wypadałoby uznać, przyjmując, że generatywno-rozpoznawczy model Fristona należy do teorii przyczynowych, że również za pomocą tej teorii nie udaje się wyjaśnić aspektowej formy reprezentacji. Wydaje się jednak, że konkluzja taka nie jest trafna, jeśli uwzględni się specyfikę bayesowskiego wnioskowania. Searle ma rację twierdząc, że na poziomie pobudzeń sensorycznych H_2O niczym nie różni się od wody, kiedy jednak weźmie się pod uwagę model generatywny, uczestniczący w procesie rozpoznawania obiektów, wówczas okaże się, że aspektowa forma stanu intencjonalnego (reprezentacji) nie stanowi problemu w tej koncepcji. Wystarczy założyć, że tego typu cechę posiadają reprezentacje modelu generatywnego, które z założenia nie są zwykłą generalizacją danych sensorycznych. Warto przypomnieć, że ich skuteczne działanie polega m.in. na uwzględnianiu kontekstu, w którym znajduje się agent oraz na „radzeniu" sobie z częściowo zniekształconymi danymi sensorycznymi. Innymi słowy, percept to rezultat działania złożonego procesu obliczeniowego (wnioskowania), a nie – jak sugeruje Searle – efekt prostego związku przyczynowo-skutkowego typu *bottom-up*. Ponadto, teoretycy hipotezy mózgu bayesowskiego wskazują na następującą możliwość: modele generatywno-rozpoznawcze można ze sobą łączyć w wielopoziomową strukturę, uzyskując mechanizm hierarchicznego wnioskowania, w którym hipotezy z poziomu wyższego są weryfikowane przez dane pochodzące z poziomu niższego lub obserwacje odnoszące się do stanów otoczenia oraz stanów wewnętrznych agenta. W ten sposób można próbować wyrazić wskazaną przez Searle'a aspektową formę stanów intencjonalnych za pomocą odpowiednio bogatego zbioru modeli generatywnych, w których uwzględnia się jedynie wybrane wymiary danych sensorycznych oraz mechanizmu selekcjonowania tych modeli generatywnych ze względu na dany kontekst.

Można założyć, że reprezentacje typu 'O', podobnie jak stany intencjonalne - przy odpowiednim schemacie wnioskowania bayesowskiego, w którym wykorzystuje się kodowanie predykcyjne, - mogą posiadać tzw. aspektową formę, tzn. udostępniać reprezentowane obiekty otoczenia pod pewnym względem. Ta intrygująca od strony filozoficznej cecha ma również swoje implikacje funkcjonalne, mianowicie pozwala agentowi filtrować informacje w zależności od jego potrzeb oraz stanu środowiska. Z perspektywy ograniczonych zasobów poznawczych, którymi dysponuje agent, jest to ważna i użyteczna dyspozycja.

Modus psychologiczny to kolejna cecha stanów intencjonalnych, którą Searle wyróżnia w swojej teorii, a której nie ma w teorii Fristona. Pełen stan intencjonalny S(p) składa się

z dwóch elementów: treści oraz modusu psychologicznego, który decyduje o typie stanu umysłowego, a co za tym idzie o kierunku dopasowania tego stanu (albo jest to dopasowanie umysł→świat, albo świat→umysł) oraz o jego wymiarze jakościowym, np. o emocjonalnym zabarwieniu danego stanu intencjonalnego. Przykładowo, oczekiwanie że będzie padał deszcz, może mieć negatywne (obawa przed deszczem) albo pozytywne (nadzieja na deszcz) zabarwienie emocjonalne. Łatwo zauważyć, odnosząc wskazaną charakterystykę do reprezentacji sensorycznych typu 'O', że podobieństwo między nimi a stanami intencjonalnymi jest tylko częściowe. Trudno przypisać reprezentacjom sensorycznym typu 'O' jakikolwiek emocjonalny charakter, choć – podobnie jak percepcyjne stany intencjonalne – posiadają one nakierowanie na zgodność, odnosząc się do określonych obiektów w świecie. Wątpliwe jest, żeby podmiot poznający miał do nich jakiś indywidualny stosunek, w szczególności emocjonalny właśnie. Choć istnieją zatem podobieństwa między reprezentacjami sensorycznymi typu 'O' a odpowiadającymi im określonymi stanami intencjonalnymi (w tym przypadku perceptami), to z pewnością nie da się zredukować tych drugich (stanów intencjonalnych) do tych pierwszych (reprezentacji sensorycznych typu 'O').

Ostatnim elementem, na który Searle mocno zwraca uwagę, jest holistyczny charakter treści stanów intencjonalnych. Holizm w sensie Searle'a to pogląd, że treść stanu intencjonalnego zależy w dużym stopniu od innych stanów intencjonalnych (patrz idea sieci stanów intencjonalnych). O ile na poziomie przekonań łatwo jest dostrzec tego typu zależności, to w przypadku stanów percepcyjnych nie jest to wcale takie oczywiste. Dopiero kiedy, przykładowo, uwzględni się raporty osób, które odzyskały wzrok w bardzo późnym wieku (Sacks, 1999, s. 122) lub przeanalizuje badania dotyczące wpływu kultury na percepcję sceny wzrokowej lub na postrzeganie osób innej rasy (tzw. efekt innej rasy) (Malinowska, 2016), to można zauważyć, że nawet na tak wydawałoby się podstawowym poziomie, jak rozpoznawanie obiektów, istnieje szereg złożonych zależności między stanami intencjonalnymi, w tym stanami percepcyjnymi. Rozpoznajemy obiekty w świecie i postrzegamy je w taki, a nie inny sposób, m.in. dlatego, że obiekty te są częścią złożonej sieci. Powiązane są m.in. z wcześniejszymi przeżyciami, przekonaniami, pragnieniami, lękami, nadziejami, itd.

Trudno byłoby jednoznacznie wskazać jakiś analogon holizmu znaczeniowego w propozycji Fristona. W koncepcji kodowania predykcyjnego zakłada się, co prawda, możliwość przeszukiwania przestrzeni hipotez (prawdopodobnych przyczyn danych

sensorycznych) w celu odnalezienia tej, która najbardziej pasuje do pozyskanych właśnie danych sensorycznych (prawdopodobieństwo $p(\nu|u;\theta)$ jest najwyższe), trudno jednak uznać, że zawarte w modelu generatywnym reprezentacje przyczyn ν są w jakiś istotny sposób ze sobą związane. Jak już wspomniano, w teorii kodowania predykcyjnego lub jej uogólnieniach (patrz: teoria aktywnego wnioskowania wykorzystująca zasadę minimalizacji swobodnej energii (*active inference and free energy principle*) (Korbak, 2019)) rozważa się złożone struktury, w których modele generatywno-rozpoznawcze zorganizowane są w hierarchiczny sposób, nadal jednak trudno byłoby przy ich pomocy wyrazić relacje występujące między stanami intencjonalnymi. Wydaje się, że również w tym przypadku reprezentacje typu 'O', choć rozwiązują złożony problem percepcyjny, to nie są wystarczające, by wyjaśnić wzajemne zależności między stanami intencjonalnymi, które uwzględniają nie tylko kontekst przyczynowy, ale również kulturowy.

Podsumowując powyższe rozważania należy stwierdzić, że reprezentacje o podobnych warunkach spełniania tworzą wielopoziomową hierarchię, której poszczególne piętra odwzorowują coraz bardziej złożone cechy środowiska oraz oczekiwania agenta. Można w ramach tej hierarchii wyróżnić proste przypadki, tj. obserwacje realizowane na zasadzie detekcji cech oraz przypadki złożone, czyli np. percepcyjne stany intencjonalne. Również działania intencjonalne, będące zasadniczym tematem niniejszej rozprawy, są zorganizowane hierarchicznie. Na najniższym poziomie występują działania proste, które – przypomnijmy – są pozbawione prior intencji i realizowane są zgodnie z wyuczoną rutyną. Na następnych poziomach występują działania złożone. Są to różnego rodzaju sekwencje działań prostych wzbogacone o rozmaite składniki: prior intencję, deliberację, plany, strategie decyzyjne, itp. Choć mamy tu do czynienia ze strukturami hierarchicznymi, to hierarchie działań nie są stałe i podlegają modyfikacjom ze względu na zmiany w otoczeniu lub w którymś z podsystemów, np. niekiedy w podejmowanym działaniu najważniejsze jest jego dokładne zaplanowanie, czyli wyraźne określenie działań składowych i kolejności ich realizowania. Kiedy indziej o podjęciu działania decyduje w pierwszej kolejności prior intencja, czyli skupienie się na zamiarze i zgrubnie określonym celu działania, a nie na sposobie jego osiągnięcia.

Wskazana hierarchia wykształca się stopniowo w trakcie rozwoju ontogenetycznego. O ile przejście od prostych obserwacji typu 'o' do obserwacji niezależnych od kontekstu typu 'O' można potraktować jako rozwój tego samego mechanizmu przetwarzania informacji, o tyle, trudno uznać powstawanie stanów intencjonalnych za proste

rozszerzenie reprezentacji niższego poziomu. „Wzbogacenie" niezależnej od kontekstu reprezentacji 'O' o aspektową formę, modi psychologiczne oraz holistycznie pojętą treść zdaje się wymagać nowych form przetwarzania informacji. Zasada minimalizacji błędu predykcji, wykorzystywana w koncepcji Fristona, jest najprawdopodobniej niewystarczająca, by wyjaśnić tego typu cechy (Loughlin, 2017).

Trudno określić, w jaki sposób dochodzi do utworzenia stanów intencjonalnych na podstawie reprezentacji niezależnych od kontekstu typu 'O'. W trakcie realizacji działań mamy do czynienia z bezpośrednim korzystaniem z reprezentacji, w tym – z testowaniem ich adekwatności i użyteczności. W konsekwencji, powinniśmy obserwować efekty działania procesu kształtującego i rozszerzającego sieć reprezentacji. Można przyjąć, że takie fenomeny, jak np. chęć wykonania ruchu czy poczucie sprawstwa towarzyszące działaniom intencjonalnym, współtworzą proces odpowiedzialny za tworzenie, rozszerzanie i modyfikowanie sieci stanów intencjonalnych, przyczyniając się do zmian reprezentacji na różnych poziomach ich organizacji, począwszy od niskopoziomowych dyspozycji tła, poprzez działania podstawowe, a skończywszy na przekonaniach, pragnieniach oraz planach. Przedstawione wcześniej wyniki badań psychologii intencji zdają się korespondować z hipotezą, którą w języku Searle'a można wyrazić w następujący sposób: **Główną funkcją składowych intencji w działaniu (są to: poczucie chęci wykonania ruchu (*sense of urge or being about to move*), odniesienie do docelowego obiektu lub zdarzenia (*reference forward to the goal object or event*)) oraz powiązanego z nimi poczucia sprawstwa jest rozwój sieci stanów intencjonalnych, w szczególności rozszerzanie jej o reprezentacje nowych związków przyczynowo-skutkowych, zidentyfikowanych na podstawie zrealizowanych działań**. Zasadność powyższej hipotezy potwierdzają następujące modele: (1) model Wolperta-Haggarda oraz (2) model Daniela Wegnera. Każdy z nich sformułowany jest we właściwym dla danych autorów języku, innym niż ten, w jakim opisane są, przedstawione w niniejszej pracy, modele złożonych działań intencjonalnych. Różnice te omawiam poniżej.

Pierwszy model (Wolperta-Haggarda) precyzuje relacje między celem, składowymi intencji w działaniu, planami motorycznymi oraz poczuciem sprawstwa. Natomiast model Daniela Wegnera określa warunki determinujące natężenie poczucia sprawstwa w zależności od kontekstu, w którym działanie jest realizowane. Obydwa modele zawierają charakterystyki specyficznych reprezentacji towarzyszących działaniu intencjonalnemu (m.in. poczucie chęci wykonania ruchu, odniesienie do docelowego obiektu lub zdarzenia),

które, choć bezpośrednio nie wpływają na jego przebieg, to współtworzą wraz z innymi stanami intencjonalnymi (m.in. pragnieniami i przekonaniami) proces racjonalizacji i interpretacji uzyskanych efektów działania.

Daniel Wegner na podstawie stworzonego modelu wyciągnął wniosek, że świadoma wola traktowana przez filozofów jako byt wyjaśniający pierwszego rzędu to konstruowana przez umysł iluzja. W jego opinii istnieje tylko sprawstwo, szczególnego rodzaju przeżycie fenomenalne (przez Wegnera definiowane jako marker somatyczny[134]) towarzyszące procesom (na ogół nieświadomym) determinującym działania. Te bardzo kontrowersyjne wnioski można przynajmniej do pewnego stopnia ograniczyć, gdy porzuci się perspektywę subiektywnego doświadczenia, w której towarzyszące działaniu zjawiska umysłowe, takie jak poczucie sprawstwa, mają je determinować lub kontrolować. W niniejszej pracy proponuje się, aby zastąpić ją perspektywą uczenia się. Proces optymalizacji i tworzenia nowych reprezentacji, który jest typowy dla zjawiska uczenia się, a który wpływa również na mechanizm kontroli zachowań agenta, wyjaśnia działanie intencjonalne bez potrzeby uznawania takich stanów umysłowych, jak intencja w działaniu czy poczucie sprawstwa – za przyczyny działania. Chciałbym w celu wykazania, że zaproponowana perspektywa jest uprawniona, pokazać najpierw, że stany umysłowe towarzyszące działaniom intencjonalnym, zidentyfikowane przez psychologów intencji, można potraktować jako stany „nadbudowane" nad wybranymi reprezentacjami występującymi w modelu 2.0. W poniższym wykazie, w lewej kolumnie zestawione są reprezentacje umysłowe wskazane zarówno przez Patrika Haggarda, jak i Daniela Wegnera, a w prawej „znaturalizowane" reprezentacje występujące w modelu 2.0:

Stany (reprezentacje) umysłowe towarzyszące działaniu intencjonalnemu	**Reprezentacje występujące w modelu działań intencjonalnych 2.0**
Haggard: „poczucie chęci wykonania ruchu" (*sense of urge*)	Poczucie chęci wykonania ruchu to stan umysłowy bezpośrednio związany z zachowaniem. W modelu zachowanie reprezentowane jest przez dwa rodzaje symboli z_t oraz Z_t. Pierwszemu odpowiada

[134] „Świadoma wola jest somatycznym markerem powiązanym ze sprawstwem, emocją, która uwierzytelnia właściciela działania jako działającego." („*Conscious will is the somatic marker of personal authorship, an emotion that authenticates the action's owner as the self."*) (Wegner, 2002, s. 327).

	program motoryczny opracowany dla określonej części ciała, wykorzystywany przez algorytm TDRL. Drugiemu – tzw. zachowaniu wysokiego poziomu – odpowiada sekwencja zachowań elementarnych (tzw. opcja – pojęcie należące do jednego z rozszerzeń metody uczenia się ze wzmacnianiem – patrz: rozdział 3.).
Haggard: „odniesienie do docelowego obiektu lub zdarzenia" (*reference forward to the goal object or event*), czyli tzw. druga składowa intencji w działaniu	Odniesienie do docelowego obiektu lub zdarzenia można powiązać z dwoma typami reprezentacji w ZMDI: • O_{t+1}, s_{t+1} – reprezentacją obserwacji lub stanu świata, do którego agent przejdzie w kolejnym kroku działania algorytmu TDRL, wykorzystując do tego wypracowaną wcześniej strategię π lub tzw. tryb eksploracji, oparty na losowym wyborze działania z_t; • r – reprezentacją nagrody, którą agent pragnie pozyskać.
Wegner: poczucie sprawstwa będące efektem nieświadomego wnioskowania, próbującego potwierdzić, że agent jest źródłem działania. Wnioskowanie bazuje na szeregu niejawnych przekonań z szeroko pojętej psychologii i fizyki ludowej (*folk psychology and folk physics).*	Zgodnie z koncepcją Wegnera poczucie sprawstwa, to stan umysłu towarzyszący aktualnie wykonanemu działaniu, wiążący działanie z zaobserwowanym stanem świata poprzez określone wnioskowanie obejmujące szereg przekonań dotyczących funkcjonowania agenta oraz środowiska. Ten złożony proces, którego zwieńczeniem jest stan poczucia sprawstwa, można

	odnieść do następujących reprezentacji obecnych w modelu: • do stanu środowiska s_t oraz s_{t+1}, • do zachowania z_t/ Z_t oraz nagrody r_{t+1} / R_{t+1} uzyskanych w wyniku przejścia od s_t do s_{t+1} (warto dodać, że warunkiem rozpoznania stanu s_{t+1} jest dokonanie obserwacji o_{t+1}/O_{t+1} na podstawie bodźców b_{t+1}); • inne reprezentacje zawarte w ZMDI, a powiązane z poczuciem sprawstwa, to zbiór dyspozycji tła (Z) (przechowywany jako sekwencje zachowań elementarnych nadzorowane przez podsystem P-H-RL-OD) oraz sieć stanów intencjonalnych zawarta w podsystemie P-ZSSI umożliwiająca przeprowadzenie nieświadomego wnioskowania.

Tym, co charakteryzuje powyższy wykaz, jest jego niejednorodność. Obiekty wymienione w lewej kolumnie tabeli to reprezentacje umysłowe ujęte z perspektywy psychologicznej, natomiast obiekty w prawej kolumnie tabeli to reprezentacje ujęte z perspektywy obliczeniowo-kognitywistycznej. Reprezentacje oznaczone symbolami b_{t+1}, o_{t+1}, O_{t+1}, z_t/Z_t, r_t / R_t, s_t, s_{t+1} mają charakter niskopoziomowy i na ogół funkcjonują na obrzeżach pola świadomości. Z kolei stany intencjonalne to reprezentacje pojęciowe, które potencjalnie mogą być dostępne w formie świadomych treści mentalnych. To na nich w głównej mierze opierają się wszelkiego rodzaju wyjaśnienia dotyczące motywów działania. Wymienione grupy reprezentacji pojawiają się na różnych etapach rozwoju osobniczego.

Pierwsza grupa reprezentacji, a przynajmniej mechanizmy odpowiedzialne za jej utworzenie, dostępne są zaraz po urodzeniu (a być może nawet w określonych fazach życia płodowego), druga grupa wymaga wielu lat uczenia się w złożonym środowisku społecznym. Fakt ten sugeruje, że b, o/O, z/Z, r/R, s mają charakter bazowy. Stany intencjonalne można z kolei zinterpretować jako reprezentacje pochodne, utworzone przez złożenie lub przekształcenie reprezentacji bazowych.

Można zaproponować, korzystając z teorii intencjonalności Searle'a, następującą interpretację: łącznie reprezentacje b_{t+1}, o_{t+1}/ O_{t+1}, z_t/Z_t, r_{t+1} / R_{t+1}, s_t, s_{t+1} tworzą strukturę, która niesie informację o elementarnych korelacjach oraz o związkach przyczynowo-skutkowych istniejących w świecie. Jej wystąpienie prowadzi do konstrukcji lub aktualizacji powiązanych z tymi reprezentacjami stanów intencjonalnych, m.in. do przekonań postaci: realizacja zachowania 'z_t/Z_t' w stanie s_t powoduje pojawienie się bodźców b_{t+1} pozwalających „skonstruować" obserwację 'o_{t+1}/O_{t+1}' oraz pozyskać nagrodę 'r_{t+1}/R_{t+1}', a także prowadzi do znalezienia się w stanie świata 's_{t+1}'. Tego typu strukturę na poziomie informacyjnym można uznać za analogon pojedynczego warunku spełniania, który zgodnie z ujęciem Searle'a, wraz z innymi tego typu warunkami, współkonstytuuje wybrane stany intencjonalne, w szczególności ich odniesienie przedmiotowe (treść).

Przedstawiona interpretacja wpisuje się również w zaproponowane przez Searle'a wyjaśnienie, które wskazuje manipulację obiektami jako metodę umożliwiającą odkrywanie związków przyczynowych, w tym również związków niezależnych od naszych działań (np. spadający z pewnej wysokości wazon ulegnie rozbiciu). Warto przypomnieć, że zdolność do rozpoznawania związków przyczynowych w świecie, zdaniem Searle'a, kształtuje się w dużej mierze na podstawie traktowania własnych zachowań (cielesnych ruchów) jako wywołanych bezpośrednio przez żywione intencje. To pojmowanie zamiaru jako siły sprawczej wywołującej cielesny ruch własnego ciała legło u podstaw wytworzenia się pojęcia przyczynowości intencjonalnej, które to pojęcie rozwija się w człowieku od najmłodszych lat. Patryk Haggard podobnie postrzega to zagadnienie (Haggard, 2005), twierdząc, że składowe intencji w działaniu wspierają proces uczenia się i umożliwiają zapamiętywanie rezultatów zachowań, aby na tej podstawie skuteczniej kontrolować działania w przyszłości. Warto zauważyć, że stwierdzenie Haggarda wzbogaca zbiór wymienionych reprezentacji o element wartościujący: wykorzystujemy składowe intencji do tego, by w przyszłości uniknąć niepożądanego (bo nieskutecznego) działania albo powtórzyć to, które okazało się korzystne dla agenta. Dlatego też na diagramie 10

przedstawiającym model 2.0 między podsystemem hierarchicznego uczenia się ze wzmacnianiem (P-H-RL-OD) a podsystemem zarządzania siecią stanów intencjonalnych (P-ZSSI) pojawia się reprezentacja nagrody (r/R) dostarczająca informacji wartościujących do (P-ZSSI) odnośnie skutków zrealizowanego zachowania, np. reakcji stresowej związanej z kolizją, do której doprowadził błędny manewr podczas prowadzenia samochodu.

Interpretacja traktująca reprezentacje b_{t+1}, o_{t+1} / O_{t+1}, z_t / Z_t, r_{t+1} / R_{t+1}, s_t, s_{t+1} jako bazowe, tj. możliwe do utworzenia i modyfikowania od urodzenia, natomiast reprezentacje stanów intencjonalnych jako nadbudowane na reprezentacjach bazowych, wpisuje się w schemat rozwoju ontogenetycznego człowieka. Uzyskujemy, traktując zależności między reprezentacjami bazowymi jako elementarne warunki spełniania, podstawę (zgodną z tym, co Searle zakłada w teorii intencjonalności) dla konstrukcji stanów intencjonalnych. Obecnie trudno jest precyzyjnie określić, jaka jest dynamika podsystemu zarządzania siecią stanów intencjonalnych, jak podsystem ten przechodzi od stanu początkowego do działania w pełnym zakresie. Wydaje się jednak, że istnieją powody, by sądzić, iż oprócz odpowiednio rozwiniętych funkcji wykonawczych (uwaga, pamięć robocza, percepcja, planowanie, monitorowanie, poznawcza elastyczność), warunkiem pełnego rozwinięcia się sieci stanów intencjonalnych jest język rozumiany jako narzędzie/mechanizm organizowania reprezentacji bazowych w reprezentacje wyższego poziomu. W opinii Searle'a, intencjonalność przysługująca wyrażeniom językowym ma charakter wtórny w stosunku do intencjonalności stanów umysłowych. Opanowanie języka wyposaża nas w zdolność do komunikowania własnych oraz cudzych intencjonalnych stanów umysłowych, a także pozwala nie tylko wyrażać wcześniej nabyte doświadczenia, ale również włączać nieznane podmiotowi przekonania, pragnienia, lęki, nadzieje, itd. w istniejącą sieć stanów intencjonalnych. Agent może w ten sposób rozszerzać sieć stanów intencjonalnych bez konieczności powtarzania związanych z tego typu reprezentacjami działań, doświadczeń czy innych form interakcji ze środowiskiem. Przykładami potwierdzającymi wpływ języka na sieci stanów intencjonalnych są wyniki badań dotyczące bliźniąt pozbawionych dostępu do języka (Luriĭa, 1959; Shaffer, 2010). Badania te pokazują, że nauka języka w istotny sposób wzbogaca dyspozycje poznawcze dziecka, rozwijając nie tylko słownictwo, ale również funkcje kognitywne oraz kontrolne. We wprowadzeniu do swojej słynnej książki *Speech and development of mental processes in the child* Łuria stwierdza:

> *Z wielu różnych źródeł wywodzi się idea, hipoteza, że znaczenie języka dla gatunku ludzkiego polega nie tyle na tym, że jest on środkiem, za pomocą którego współpracujemy i komunikujemy się ze sobą, ile na tym, że umożliwia każdemu z nas – czy to jako jednostce, czy jako części pewnej zbiorowości - reprezentowanie świata w takiej postaci, w jakiej go doświadczamy: i tym samym konstruujemy - chwila po chwili, rok po roku – jego całościowy obraz.*[135]

Sformułowana w przywołanym cytacie hipoteza Łurii koresponduje z teorią intencjonalności Searle'a i wskazuje język (oraz związane z nim mechanizmy) jako ważne narzędzie służące do odwzorowywania świata za pomocą reprezentacji umysłowych (stanów intencjonalnych). Zdolność do wyrażania w formie językowych stwierdzeń związków między intencją w działaniu a zaobserwowanym efektem nie tylko pozwala nam komunikować ich treść innym osobom, ale przede wszystkim pozwala włączać je w rozumowania czy inne operacje mentalne. Operacje te nie byłyby możliwe bez nadania im językowej formy, gdyż zaangażowana w nie liczba reprezentacji przekracza bazową pojemność pamięci roboczej. W tym kontekście rozszerzająca się za sprawą języka sieć stanów intencjonalnych może być pojmowana jako odpowiednio zorganizowana baza wiedzy, w której poszczególne elementy aktywuje się „na żądanie", w zależności od kontekstu wyznaczanego przez konstruowany plan działania.

Język, oprócz dostępu do „bazy wiedzy", umożliwia również identyfikowanie podobieństw formalnych między poszczególnymi reprezentacjami. Wskazana przez Searle'a aspektowa forma stanów intencjonalnych, zakładająca ich niejawną kategoryzację – w połączeniu z językowymi mechanizmami odpowiedzialnymi za konstruowanie pojęć – umożliwia dostrzeganie nietrywialnych związków pojawiających się między reprezentacjami a związanymi z nimi stanami świata. Można zatem mówić o swoistym sprzężeniu zwrotnym, które pojawia się między podsystemami kontroli zachowań a językowo zorganizowaną siecią stanów intencjonalnych. Z jednej strony, niskopoziomowe reprezentacje, które towarzyszą zachowaniom, przekształcane są za pomocą języka w zorganizowaną wiedzę dostępną w formie sieci stanów intencjonalnych (m.in. dobrze uzasadnionych przekonań), a z drugiej, uzyskana w ten sposób wiedza wykorzystywana

[135] "From many diverse sources has come the idea, the hypothesis, that the importance of language to mankind lies not so much in the fact that it is the means by which we cooperate and communicate with each other as in the fact that it enables each of us, as individuals and in cooperation, to represent the world to ourselves as we encounter it: and so to construct – moment by moment and year after year – a cumulative representation of 'the world as I have known it". (Luriia, 1959, s. 7).

jest do planowania nowych celów i działań. Warto zauważyć, że wskazane mechanizmy istotnie wykraczają poza funkcjonalne możliwości podstawowego mechanizmu uczenia się ze wzmacnianiem opierającego się w głównej mierze na informacji wartościującej (tj. na nagrodach), który to mechanizm uniemożliwia zaawansowane planowanie.

Do tej pory nie udało się ustalić, jak przebiega proces przekształcania reprezentacji elementarnych w uchwytne językowo stany intencjonalne. Nie wiadomo, jak reprezentacje z_t, b_t i O_t, O_{t+1} są kategoryzowane i włączane albo usuwane ze struktury stanu intencjonalnego ani skąd bierze się ich aspektowa forma. Nieznany jest również sposób, za pomocą którego łączą się one w złożoną sieć, tworząc nową, wysokopoziomową treść. Wiadomo natomiast, że przynajmniej do pewnego stopnia operacje przekształcające reprezentacje (b_{t+1}, o_{t+1} / O_{t+1}, z_t / Z_t, r_{t+1} / R_{t+1}, s_t, s_{t+1}) mają charakter interpretacyjny i konstruktywistyczny. Tego typu wniosek wyprowadzić można z modelu Daniela Wegnera (patrz: rozdział 4. Podpunkt: *Post-rekonstruktywistyczny model sprawstwa Daniela Wegnera*), który dostrzegł w przeżyciu sprawstwa (*agency*) przejaw złożonego procesu poznawczo-emocjonalnego. Podstawowym zadaniem procesu odpowiedzialnego za konstrukcję sprawstwa jest zinterpretowanie uzyskanych rezultatów działania w taki sposób, by było to spójne z dotychczas zgromadzonymi doświadczeniami oraz wiedzą podmiotu, a także – żeby pozwalało wyjaśnić w kategoriach mentalnej przyczynowości, dlaczego dane działanie zostało zrealizowane. Złożoność tej interpretacji może być mniejsza lub większa w zależności od etapu rozwoju ontogenetycznego oraz od okoliczności towarzyszących działaniom intencjonalnym. „Wrażliwość" poczucia sprawstwa na kontekst, w którym jest ono konstruowane, świadczy o występowaniu procesów, które każdorazowo podczas realizacji działania rozwiązują pewne fundamentalne zadania poznawcze dające się wyartykułować w formie pytania: „czy przyczyną zaobserwowanego przeze mnie stanu świata było moje działanie intencjonalne, czy też stan ten spowodowany został bez mojego udziału?". Ludzki umysł, rozwiązując tego typu „zagadkę", na bieżąco aktualizuje, powiększa i wzbogaca swoją bazę przekonań. Im wyraźniejsze poczucie sprawstwa, tym silniejsze przekonanie dotyczące przyczyn oraz skutków zachowania. Im słabsze poczucie sprawstwa, tym słabsze (mniej prawdopodobne) przekonanie o wpływie agenta na wywołaną działaniem zmianę w świecie.

Przedstawiony proces towarzyszący działaniom intencjonalnym, który powoduje rozszerzanie i kształtowanie sieci stanów intencjonalnych, rzadko pojawia się w kontekście rozważań dotyczących kontroli zachowań. W przedstawionej propozycji powiązanie

zachowań i stanów intencjonalnych ma jednoznaczny i dobrze zdefiniowany charakter. Niskopoziomowe reprezentacje wykorzystywane przez mechanizm uczenia się ze wzmacnianiem stanowią „wejście”[136] dla procesów poznawczych, za pomocą których można skonstruować reprezentacje wyższego poziomu oraz włączyć je w złożoną sieć stanów intencjonalnych.

Przedstawiony model 2.0 (i objaśnienia jego działania) trudno uznać za kompletny model działania intencjonalnego, choć trafniej niż wersja 1.2 odtwarza jego składniki oraz ich funkcje. Nie jest możliwe opisanie tak złożonej struktury, jaką jest sieć stanów intencjonalnych w kilku zwięzłych paragrafach, można jedynie zasygnalizować jej złożoność, a także wskazać dalsze kierunki rozwijania modelu 2.0. Z perspektywy zaproponowanego tu zintegrowanego modelu działań intencjonalnych najważniejszy jest następujący wniosek: podsystem uczenia się ze wzmacnianiem „zasila” elementarnymi reprezentacjami podsystem zarządzania siecią stanów intencjonalnych, ta z kolei stanowi podstawę procesów deliberacji i planowania.

Dotychczas omówione modele pomijały ważny aspekt złożonych działań intencjonalnych, a mianowicie, że ich inicjowanie, a następnie wykonywanie jest stosowaniem się do uprzednio wypracowanego planu. Kolejny model działania intencjonalnego (3.0) został wzbogacony o podsystem odpowiedzialny za planowanie działania. Podsystem planowania – oznaczony symbolicznie jako (P-PRP) – pełni w przyjętym rozwiązaniu rolę swoistego „translatora” stanów intencjonalnych na zachowania. Podsystem ten współpracuje zarówno z podsystemem zarządzającym siecią stanów intencjonalnych, jak i z podsystemem hierarchicznego uczenia się ze wzmacnianiem.

5.3.5 Model 3.0 – działanie intencjonalne realizowane według planu

Rozważania prowadzone w poprzednim podrozdziale dotyczyły w głównej mierze mechanizmów związanych z rozbudową sieci stanów intencjonalnych, struktury przechowującej złożone reprezentacje otaczającego świata, które odwzorowują rzeczywistość w znacznie bardziej abstrakcyjny i całościowy sposób, niż metoda uczenia

[136] Określenie „wejście” (input) użyte zostało zgodnie z funkcjonującą w informatyce konwencją. Konwencja ta traktuje informacje lub polecenia dopływające do systemu jako to tzw. dane wejściowe. Przykładami tego typu danych są dane początkowe, inicjujące działanie systemu, dane przesyłane do systemu z zewnątrz, itd.

się ze wzmacnianiem. W ten sposób w modelu 2.0 stworzone zostały fundamenty dla podsystemu planowania, którego głównym zadaniem jest zwiększenie efektywności agenta poprzez włączenie wiedzy dziedzinowej w proces osiągania celu. Model 3.0 działania intencjonalnego jest zatem wzbogacony w porównaniu z modelem 2.0 o podsystem planowania i realizacji planów (P-PRP).

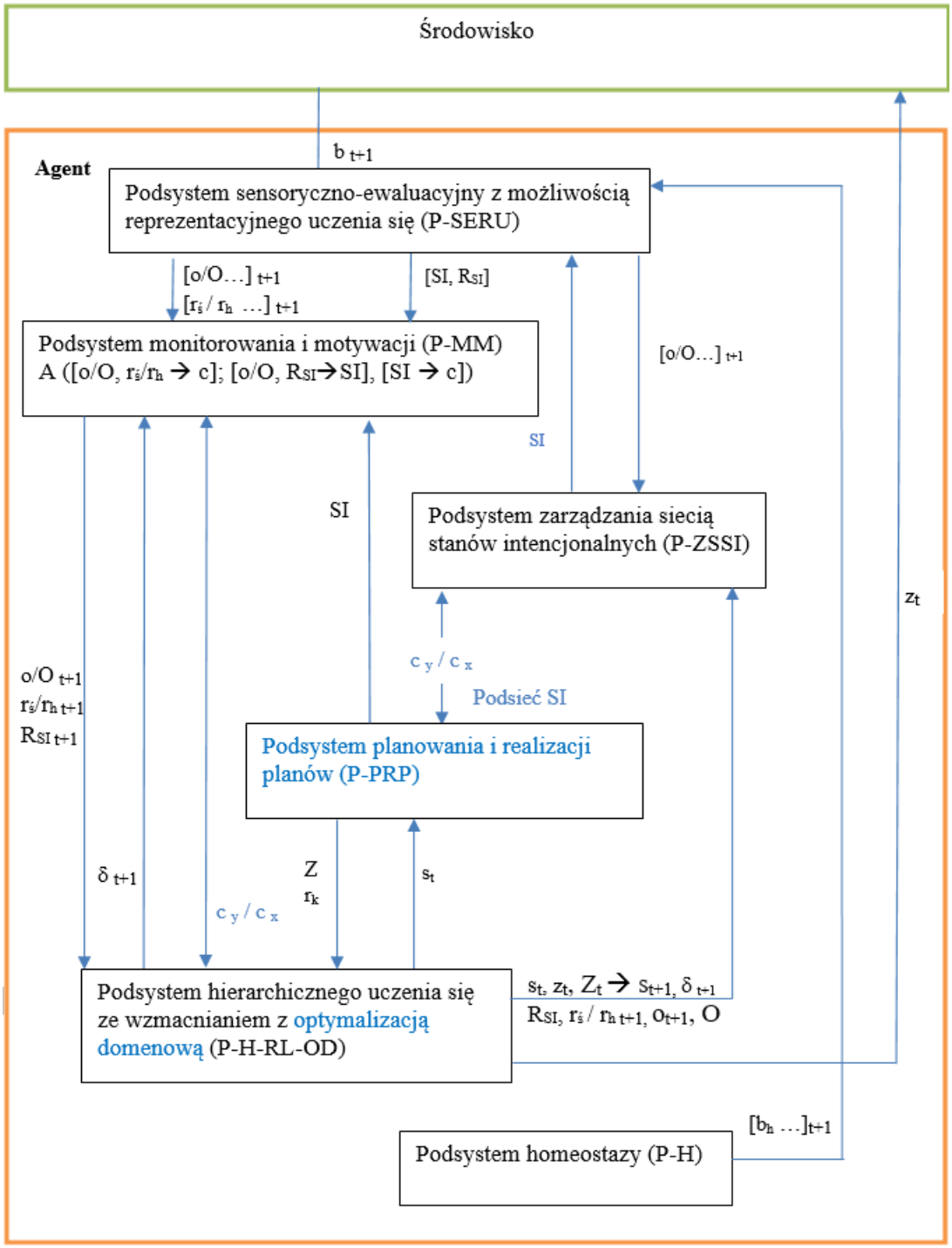


Diagram 14. Model działania intencjonalnego z podsystemem zarządzania siecią stanów intencjonalnych (P-ZSSI).

Legenda symboli (uzupełnienie w odniesieniu do Modelu 2.0):

- r_k – nagrody kształtujące służące do przekazywania wiedzy domenowej do podsystemu P-H-RL-OD, istotnej z perspektywy realizowanego planu.

Główna funkcja podsystemu P-PRP wymaga doprecyzowania, w szczególności wyjaśnić należy mechanizm konstrukcji planów oraz sposób, w jaki odbywa się współpraca starszego ewolucyjnie podsystemu uczenia się ze wzmacnianiem z podsystemem planowania. W związku z tym proponuję następujące hipotezy:

H1: Plan jest efektem współpracy podsystemu zarządzania siecią stanów intencjonalnych oraz podsystemu planowania i realizacji planów.

Postulowana w hipotezie H1 współpraca pomiędzy dwoma podsystemami polega na tym, że ten pierwszy (P-ZSSI) dostarcza różnego rodzaju modele (środowiska, agentów, zdarzeń w świecie), na podstawie których drugi podsystem (P-PRP) opracowuje – na bazie dostępnych reprezentacji – wysokopoziomowe scenariusze realizacji celu, przeprowadza ich ewaluację i wybiera najbardziej korzystny z nich. Tak skonstruowany plan ma charakter wstępny i prowizoryczny. Jego aktywacja odbywa się za pośrednictwem prior intencji „zainstalowanej" w podsystemie monitorowania i motywacji (P-MM).

H2: Wraz z rozwojem osobniczym do niskopoziomowej kontroli zachowań dołączona zostaje kontrola wysokopoziomowa.

Na niższym poziomie dobór zachowań kontrolowany jest przez podsystem hierarchicznego uczenia się ze wzmacnianiem (P-H-RL-OD), natomiast na wyższym poziomie dobór poszczególnych działań wykonywany jest na podstawie prowizorycznego planu kontrolowanego przez podsystem realizacji planów (P-PRP).

Hipotezy H1 i H2 odnoszą się wprost do dwóch głównych funkcji podsystemu P-PRP: (1) do konstrukcji planów oraz do ich (2) realizacji. Wspólnie charakteryzują one mechanizm kontroli zachowań, wykraczający poza możliwości metody uczenia się ze wzmacnianiem (nawet w wersji zhierarchizowanej). Uzasadnieniem dla hipotezy pierwszej (H1) są przeanalizowane w pracy propozycje teoretyczne, przede wszystkim są to: (i) badania z obszaru uczenia maszynowego (patrz: opisane w rozdziale trzecim rozszerzenia metody uczenia się ze wzmacnianiem), (ii) schemat przebiegu działań intencjonalnych Searle'a (wraz z modelem Mialla i Wolperta (1996)) oraz (iii) krytyczna analiza Montague, która dotyczy kosztów i ograniczeń szczegółowego planowania.

Pierwszy z wymienionych obszarów – uczenie maszynowe – jednoznacznie i precyzyjnie definiuje problem, który próbuje się rozwiązać za pomocą planowania. Skalowalność algorytmu, tzn. możliwość stosowania go w złożonym środowisku, zgodnie z przedstawionymi analizami (patrz: punkt 3.4 w rozdziale trzecim), jest głównym problemem podstawowej wersji metody uczenia się ze wzmacnianiem. Okazuje się, że im bardziej złożona definicja stanu świata, im większa przestrzeń, w której funkcjonuje agent, tym dłuższy – w ogólności nawet wykładniczy – czas znajdowania optymalnej strategii doboru zachowań. W związku z tym, metoda uczenia się ze wzmacnianiem, bez dodatkowych rozszerzeń, nie pozwala na zakończenie procesu eksploracji w akceptowalnym czasie, a tym samym nie zapewnia agentowi odpowiednio efektywnej strategii zachowań. Dlatego też zaczęto poszukiwać rozszerzeń algorytmów RL, które pozwoliłyby przezwyciężyć, a przynajmniej ograniczyć problem długotrwałej eksploracji. Do najważniejszych tego typu rozszerzeń należą: funkcje aproksymujące, uwzględnienie zachowań wyższego rzędu (tzw. opcji) oraz dodanie zewnętrznego mechanizmu planowania opartego na proceduralnej wiedzy domenowej.

Z perspektywy zintegrowanego modelu działań intencjonalnych oraz hipotezy H1 najważniejszy wniosek płynący z przedstawionych badań jest następujący: plan nie musi być kompletny i precyzyjny, by skutecznie wspierać realizację celów. Jego zadanie nie polega na wyznaczeniu precyzyjnej sekwencji zachowań, a jedynie na ograniczeniu zakresu eksploracji. Plan stanowi rodzaj zbioru wskazówek dla podsystemu realizującego cele. W tej interpretacji, wzbogacenie struktury modelu 2.0 o podsystem, który tworzy zgrubny plan postępowania, jest istotnym przybliżeniem do faktycznych mechanizmów działania intencjonalnego.

Analiza Searle’a, a w szczególności zaproponowany przez niego schemat działania intencjonalnego (patrz: Schemat przebiegu działania intencjonalnego wg Searle’a przedstawiony w rozdziale 2.) jest drugą, istotną składową hipotezy dotyczącej źródeł powstawania planów. Można, wykorzystując ten schemat, wstępnie określić zasadę działania procesów odpowiedzialnych za tworzenie planów oraz doprecyzować typy kształtujących je reprezentacji. Warto przypomnieć, że amerykański filozof wyróżnił w schemacie działania intencjonalnego trzy fazy: (1) fazę deliberacji, (2) fazę inicjowania działania oraz (3) fazę warunkowej kontynuacji. Z perspektywy planowania kluczowa jest pierwsza faza, podczas której agent rozważa możliwe scenariusze prowadzące do osiągnięcia celu. Końcowym rezultatem fazy pierwszej jest prior intencja, czyli stan

intencjonalny, którego warunki spełniania określają, jakie działania są niezbędne do tego, aby dany scenariusz został zrealizowany. Tak określoną intencję Patryk Haggard wyraża poprzez formułę odnoszącą się do procesów przekształcania informacji. Jego zdaniem intencja to

> *[…] zbiór powiązanych ze sobą procesów informacyjnych, odpowiedzialnych za przekształcanie pragnień i celów w zachowanie.* (patrz: rozszerzony model Mialla i Wolperta (Haggard, 2005, s. 290)).

Określenie Haggarda pozwala ująć prior intencję zarówno jako reprezentację jak i jako proces obliczeniowy. W ten sposób Haggard trafnie oddaje szczególny status prior intencji, czyniąc z niej swoisty „pomost" prowadzący do ważnych składowych celu: (1) decyzji o jego realizacji (intencja, inaczej niż pragnienia, zawiera, zdaniem Searle'a, tzw. samo-odniesienie przyczynowe, czyli zdolność do kauzalnego powodowania działań) oraz (2) specyfikacji najważniejszych elementów planu, które są niezbędne, by osiągnąć cel. Warto przypomnieć, że konstrukcja prior intencji w fazie deliberacji uruchamia szereg procesów poznawczych, które m.in. konstruują i modelują możliwe scenariusze realizacji celu, szacują prawdopodobieństwa zakończenia ich sukcesem, dokonują predykcji oraz ewaluacji końcowych wyników, itd. Dla tego typu działań nieodzowne są odpowiednio rozwinięte dyspozycje poznawcze, czyli zdolność do wykonywania złożonych rozumowań (logicznych, abdukcyjnych, heurystycznych), wyobrażanie sobie przyszłych stanów świata na podstawie wcześniejszych doświadczeń lub zdobytej wiedzy teoretycznej czy wreszcie przełączanie między różnymi strategiami rozwiązywania problemów. Bez wymienionych procesów planowanie staje się niemal niemożliwe, dlatego tak trudno dostrzec jego przejawy w świecie zwierząt.

Dokładniejsza analiza wymienionych zagadnień wykracza, ze względu na ich złożoność, poza ramy niniejszej pracy. Jednak nawet zaproponowana tu ich zgrubna specyfikacja pokazuje, dlaczego konstrukcja planów w modelu 3.0 jest wynikiem współpracy dwóch podsystemów. Przyjęta architektura – zaprezentowana na diagramie 3.0 oraz sformułowana za pomocą hipotezy H1 – jest rezultatem połączenia dwóch podejść: (1) holizmu treściowego zawartego w teorii intencjonalności oraz (2) poszukiwań badawczych, których celem jest znalezienie korelatów neuronalnych dla wybranych dyspozycji poznawczych (np. rozumowań (Goel i in., 1998)). Jeśli reprezentacje wchodzące w skład sieci stanów intencjonalnych mają cechować się holistyczną treścią, to naturalne wydaje się rozwiązanie oparte na pojedynczym podsystemie, który zapewni, że

dodawanie do sieci nowych węzłów, modyfikowanie istniejących czy usuwanie nieaktualnych będzie zarządzane przez pojedynczy podsystem uspójniający całą sieć. Wyniki badań nad neuronalnymi korelatami wybranych funkcji mózgu-umysłu wskazują, że wiele z tych domniemanych korelatów posiada specyficzną lokalizację w mózgu. Gdy w jakiejś sytuacji dojdzie do uszkodzenia wybranego obszaru mózgu (pojawi się lezja), wówczas można obserwować pojawienie się jakiegoś deficytu, np. pacjent nie potrafi posługiwać się pojęciami abstrakcyjnymi albo błędnie szacuje ryzyko, albo nie potrafi wyhamować określonych reakcji, albo nie rozpoznaje twarzy znajomego, albo cierpi na innego typu zespół. W modelu 3.0 przyjmuje się, że podsystem zarządzania siecią stanów intencjonalnych ma uniwersalny charakter, a także jest on szczególnego rodzaju repozytorium przechowującym złożoną sieć przekonań, pragnień, celów, intencji, obaw, itd. Od strony funkcjonalnej jego zadanie sprowadza się do tego, by w trybie ciągłym aktualizować stan sieci oraz udostępniać innym podsystemom – na żądanie lub w formie „notyfikacji" – fragmenty (podsieci) lub wybrane stany intencjonalne, dlatego na diagramie przedstawione są trzy ważne połączenia istniejące między podsystemem zarządzania siecią stanów intencjonalnych (P-ZSSI) a następującymi podsystemami:

- (P-SERU) sensoryczno-ewaluacyjnym z możliwością reprezentacyjnego uczenia się – tego typu związek umożliwia funkcjonowanie modeli generatywno-rozpoznawczych postulowanych przez Karla Fristona (Friston, 2010), „wrażliwych" m.in. na system przekonań i pragnień podmiotu;
- (P-PRP) planowania i realizacji planów – związek ten umożliwia konstrukcję planów oraz ich aktywowanie za pomocą prior intencji;
- (P-H-RL-OD) hierarchicznego uczenia się ze wzmacnianiem z optymalizacją domenową – wskazana relacja odpowiada za dostarczanie do podsystemu P-ZSSI reprezentacji z podsystemu P-H-RL-OD zorganizowanych w następujący związek: $(s_t, z_t, Z_t \rightarrow s_{t+1}, \delta_{t+1}, R_{SI}, r_{t+1}, o_{t+1}, O_{t+1})$ – symbole przed strzałką odnoszą się do stanu świata przed wykonaniem zachowania, natomiast symbole za strzałką – to reprezentacje dostępne po jego wykonaniu. Tego typu związek ma duży wpływ na rozwój sieci stanów intencjonalnych na jej początkowym etapie.

Sieć stanów intencjonalnych w modelu 3.0 pełni funkcję „usługową" dla dwóch pierwszych podsystemów, stanowiąc rodzaj samozarządzalnego repozytorium, z którego wyspecjalizowane procesy mogą czytać informacje lub zapisywać rezultaty swojego

działania. Tak dzieje się na przykład w przypadku prior intencji oraz planów, które z jednej strony funkcjonują jako reprezentacje podsystemu planowania, a z drugiej – jako element sieci stanów intencjonalnych. Zaproponowane podejście zakłada, że dostępne w podsystemie P-ZSSI dane mogą być współdzielone, tzn. stan wybranej reprezentacji może być odczytany lub zmodyfikowany przez więcej niż jeden podsystem. Dobrym przykładem jest tu prior intencja, gdyż jest ona włączona do podsystemu motywacji i monitorowania (P-MM), a zarazem funkcjonuje jako element podsystemu planowania i realizacji planów (P-PRP).

Gdy rozważa się uzasadnienie hipotez H1 i H2 dotyczących źródeł i mechanizmów konstruowania planów, to nie można pominąć wyników badań przeprowadzonych przez Reada Montague. Ten amerykański neuronaukowiec, wyjaśnia niską efektywność drobiazgowego planowania, proponując wyobrażenie sobie następującego zadania:

> *Przypuśćmy, że chcę zbudować robota, który będzie umiał bezpiecznie łowić ryby na Alasce. Jak moglibyśmy zaprogramować robota, by osiągnąć ten cel? Moglibyśmy spróbować stworzyć złożony algorytm podejmowania decyzji. Tego typu projekt wymagałby starannego* ***planowania*** *i olbrzymiej* ***wiedzy*** *na temat trudności, jakie robot mógłby napotkać na Alasce, musielibyśmy się też bardzo wysilić, by przewidzieć wszystkie możliwe niebezpieczeństwa i niespodzianki. W pewnym sensie natknęlibyśmy się na te same problemy, z którymi trzeba się było mierzyć podczas wysłania łazika na powierzchnię Marsa.*[137]

Przytoczony przykład dobrze oddaje, zdaniem Montague, złożoność podjętego zagadnienia. Liczba możliwych zdarzeń, sytuacji, różnego rodzaju niebezpieczeństw jest tak duża, że próba zaimplementowania wskazanej wiedzy w formie preskryptywnego algorytmu z pewnością zakończyłaby się niepowodzeniem. Montague zwraca uwagę na jeszcze jeden, równie fundamentalny problem, mianowicie na obecną i nieusuwalną ze świata przyrody niepewność. Przejawia się ona m.in. w zachowaniach zwierząt, które poprzez nagłe i nieoczekiwane manewry utrudniają czyhającym na nie drapieżnikom możliwość przewidzenia ich reakcji. W konsekwencji, dokładne planowanie przyszłych

[137] „Suppose that I want to build a robot to achieve the goal "catch fish safely in the Alaskan wilderness." How could we possibly program this into the robot? We might rush into the problem and try to build complex decision-making algorithms around this global goal. Our robot-building project would require a lot of planning and knowledge about the difficulties the robot might encounter, and we would have to make some really good guesses about unexpected dangers and other surprises. We would encounter the same problems that faced the recent mobile probes sent to the Martian surface.” (Montague, 2006, s. 49).

zachowań z góry skazane jest na niepowodzenie i dlatego tak rzadko można je spotkać w środowisku naturalnym. W tej sytuacji, jak twierdzi Montague, jedyną sensowną alternatywą jest odpowiednio zmodyfikowana metoda uczenia się ze wzmacnianiem. Niestety, propozycja rozszerzenia algorytmu przedstawiona przez niego w *Why Choose This Book?* jest bardzo ogólna. Polega ona na dodaniu do układu selekcjonującego zachowania swoistego „przewodnika" (*„guide" signals*) zawierającego wskazówki, podpowiedzi i sugestie dotyczące tego, czego należy unikać, na co zwrócić uwagę, gdzie szukać pomocy w razie niebezpieczeństwa. Tak zorganizowana wiedza nie jest instrukcją wykorzystywaną przez agenta w celu poruszania się od punktu startowego do docelowego, ale luźnym zbiorem reguł odnoszących się do wybranych cech środowiska, w szczególności do tych, które w jakiś istotny sposób odbiegają od normy lub są niebezpieczne dla agenta. Na przykład, na Alasce należy unikać miejsc typu X, bo często można w nich spotkać niedźwiedzie. By uniknąć pogryzienia przez komary, należy zabrać ze sobą moskitierę. Tego typu wskazówki, jak twierdzi Montague, można sprowadzić do postaci sygnałów kary i nagrody, czyli kluczowej informacji wartościującej, gdy uwzględnia się perspektywę metody uczenia się ze wzmacnianiem. Niestety, propozycja amerykańskiego badacza tylko częściowo pokrywa się ze stanem badań specjalistów z obszaru uczenia maszynowego.

Można zauważyć, analizując przedstawione w rozdziale trzecim niniejszej rozprawy rozszerzenia algorytmu RL, że włączenie wiedzy domenowej w proces uczenia się ze wzmacnianiem generuje cały szereg trudności (Grześ, 2010). Najbardziej zbliżona do koncepcji Montague jest koncepcja nagród kształtujących (*shaping rewards*). To za ich pośrednictwem można włączyć w działanie algorytmów RL dodatkową, niedostępną w środowisku, informację wartościującą. Gdy na przykład robot będzie miał wykonać następujące zadanie: „odkurzyć ściśle określone miejsca w budynku złożonym z szeregu pomieszczeń i korytarzy", wówczas dodatkową wiedzę na temat środowiska, odnoszącą się np. do tego, że „w danym pomieszczeniu istnieje tylko jedno miejsce wymagające odkurzenia", można wyrazić w formie dodatkowych nagród kształtujących. W ten sposób znacząco ogranicza się zakres eksploracji, a tym samym skraca się czas niezbędny do opracowania optymalnej strategii zachowań.

Należy, biorąc w nawias wątpliwości odnoszące się do rozwiązania zaproponowanego przez Montague, uwzględnić jego argumenty wskazujące na ograniczenia planowania jako metody kontroli zachowań. W niniejszej pracy proponuje się rozwiązanie kompromisowe,

łączące element preskryptywny (planowanie prowizoryczne) z mechanizmem uczenia się ze wzmacnianiem.

Poniżej przedstawione zostaną najważniejsze racje uzasadniające przydatność prowizorycznego planowania.

Racja 1: Elementarne formy planowania polegają na sekwencjonowaniu zachowań wyższego poziomu.

Model działania intencjonalnego w wersji 1.2 rozszerzony został w stosunku do wersji 1.1 o mechanizm hierarchizacji zachowań, umożliwiający m.in. wykorzystywanie nabytych wcześniej doświadczeń w nowych okolicznościach. Najprostsza forma planowania działań staje się możliwa za sprawą zachowań wyższego rzędu, konstruowanych za pomocą mechanizmu hierarchizacji zachowań. Skuteczne opracowanie planu oraz jego realizacja w tym przypadku wymaga od agenta dysponowania dwojakiego rodzaju informacją: (1) powiązaniem (Z, r/R) zachowania wyższego rzędu ‘Z’ z typem nagrody ‘r/R’, do pozyskania której może ono zostać wykorzystane (takie powiązanie spowoduje, że ‘Z’ uzyska status celu, który w sprzyjających warunkach zostanie zrealizowany przez podsystem P-H-RL-OD) (2) powiązanie (Z, s) zachowań wyższego rzędu ‘Z’ z reprezentacją stanu świata ‘s’, do którego agent przejdzie po zrealizowaniu zachowań Z.

Pierwsze powiązanie (Z, r/R) zostało zarysowane w sekcji 5.3.3 przy okazji omawiania mechanizmów odpowiedzialnych za powstawanie zachowań wyższego rzędu. Przypomnę, że jednym z głównych warunków pojawienia się zachowania wyższego rzędu jest zdolność podsystemu hierarchicznego uczenia się do generalizacji, tj. do łączenia sekwencji zachowań niskopoziomowych w większe całości (tzw. opcje) oraz wiązania ich ze stanami świata, do których prowadzą. W przypadku organizmów o dostatecznie rozwiniętych układach nerwowych zdolność ta może działać na podobnej zasadzie, jak modele generatywno-rozpoznawcze „rozwiązujące” problem tworzenia reprezentacji wolnych od kontekstu ‘O’. Prawdopodobnie na analogicznej zasadzie funkcjonują uogólnione sekwencje zachowań elementarnych, czyli tzw. opcje ‘Z’ (w terminologii Searle’a nazywane: „działaniami podstawowymi”), które zostały utworzone na bazie powtarzających się fragmentów strategii doboru zachowań (*policy*), konstruowanej w ramach metody uczenia się ze wzmacnianiem.

Drugie z wymienionych powiązań (Z, s), łączące zachowanie wyższego rzędu z jego faktycznymi skutkami, jest przejawem poznawczych zdolności podmiotu. Podmiot ten na odpowiednim etapie rozwoju może skojarzyć reprezentację typu ‘Z’ z charakterystycznymi dla danego zachowania skutkami i przekształcić tego typu asocjację w przekonanie lub zbiór przekonań dotyczących przyczynowego wpływu agenta na otoczenie. Opisana zdolność w dużym stopniu opiera się na mechanizmie, który odpowiada za rozpoznawanie niezależnych od działań podmiotu związków przyczynowych. Przyczynowość intencjonalna, zdaniem Searle’a, która jest przez agenta doświadczana na co dzień, wraz z jego nastawieniem (*stance*) (będącym częścią „tła” (*background*)), polegającym na dostrzeganiu w świecie regularności, pozwala agentowi identyfikować relacje przyczynowe między zdarzeniami niezależnymi od niego. Sformułowany na podstawie obserwacji zdarzeń sąd: „mój rzut kamieniem spowodował rozbicie wazy” – wraz z upływem czasu zostaje uogólniony i uzyskuje postać: „poruszający się kamień o odpowiedniej masie (przyczyna), rozbija porcelanowe naczynie (skutek), niezależnie od tego, kto (lub co) zainicjował(o) jego ruch”. Podobnie traktowane są wielokrotnie powtarzane sekwencje zachowań oraz ich rezultaty. Początkowo są one utożsamiane z mechanizmem optymalizującym osiągnięcie konkretnego celu (uniwersalną umiejętnością), następnie autonomizują się i dlatego zaczynają funkcjonować jako przekonania dotyczące świata. Wiedza, jaką dysponujemy, nie pozwala na precyzyjne zrekonstruowanie mechanizmu obliczeniowego związanego z konstruowaniem tego typu przekonań. Z pewnością, wielokrotnie powtarzane sekwencje zachowań, które uzyskały status zachowania typu Z, są w jakiś sposób powiązane z Wegnerowskim pojmowaniem tego, jak przebiega proces konstrukcji poczucia sprawstwa. Proces ten, jak już wspomniano, realizuje potrzebę podmiotu odróżniania działań własnych od działań innych agentów oraz odpowiada za tworzenie przekonań odnoszących się do skutków działań. Postulowane przez Wegnera zasady: priorytetu, spójności i wyłączności dobrze nadają się do tak zdefiniowanego zadania. Wskazane przez amerykańskiego psychologa ograniczenia, które muszą być spełnione, by agent uznał się za sprawcę danego działania, z jednej strony wykorzystują zgromadzoną dotychczas w sieci stanów intencjonalnych wiedzę, a z drugiej – mogą ją modyfikować i uzupełniać.

Racja 2: Zaawansowane planowanie skorelowane jest ze złożonością środowiska, dyspozycjami poznawczymi agenta, znajomością domeny oraz potencjalnymi stratami będącymi skutkiem popełnionego błędu.

Z argumentacji Montague wnosić można, że planowanie polegające na konstrukcji szczegółowych sekwencji działań w środowisku naturalnym, o ile występuje, jest wyjątkiem od reguły. Złożoność i nieprzewidywalność środowiska naturalnego przekraczają poznawcze możliwości zwierząt, dlatego ich strategie zachowań opierają się głównie na metodzie uczenia się ze wzmacnianiem oraz na wrodzonym, odziedziczonym po przodkach, repertuarze zachowań. W przypadku *homo sapiens* sytuacja jest znacznie bardziej skomplikowana. We wszystkich znanych kulturach można zaobserwować zaawansowane formy planowania, w szczególności dotyczące podziału ról w społeczeństwie (Sztompka & Konieczny, 2005, s. 93), realizacji wspólnych przedsięwzięć (Searle, 1995b) czy praktykowania rytuałów religijnych (Harari, 2017).

Różnorodne kultury są przykładami na to, jak planowanie może wspomóc metodę uczenia się ze wzmacnianiem. Warto w tym kontekście rozważyć przypadek polowania, którego „skuteczne przeprowadzenie" wymaga odpowiedniej koordynacji działań, zwłaszcza w większej grupie myśliwych. Myśliwi znają zwyczaje danego gatunku zwierząt, nadal jednak nie potrafią przewidzieć, jak zachowa się konkretne zwierzę – czy to jeleń, czy też dzik albo niedźwiedź. W takiej sytuacji najbardziej „efektywny" może się okazać model oparty na wstępnym i prowizorycznym planie, który będzie potrzebny do wyznaczenia głównych etapów polowania, który okaże się przydatny przy wyborze najważniejszych form zachowań, a równocześnie będzie użyteczny podczas gromadzenia informacji napływających z otoczenia. Taki plan nie będzie wymagał wielkich inwestycji energetycznych, czyli długotrwałej deliberacji, a jednocześnie w znaczący sposób ograniczy zakres ewentualnych błędów i zbędnych działań. Plany szczegółowe funkcjonują niejako w opozycji do planów prowizorycznych. Ich przygotowanie wymaga żmudnych wysiłków i znacznych nakładów energetycznych. Dobrym przykładem takich szczegółowych planów, z jednej strony bardzo kosztownych, a z drugiej niezbędnych dla skutecznej realizacji tego typu przedsięwzięć, są projekty budowalne wymagające wielu przygotowań i niezwykle sumiennego wykonania. Pomimo wielu starań, niemal każdy, kto miał do czynienia z ekipami budowlanymi, wie, jak wiele błędów jest popełnianych w trakcie realizacji planu i jak trudno jest inwestorowi zapanować nad koordynacją działań poszczególnych osób zaangażowanych w projekt. Można stwierdzić, uogólniając powyższy przykład, że zastosowanie szczegółowych planów jest uzasadnione tam, gdzie – ze względu na czasochłonność i szeroki zakres realizowanego projektu – pragniemy w znaczący sposób ograniczyć liczbę nieprzewidywalnych zdarzeń w środowisku, w szczególności tych, które mogą być niebezpieczne lub niekorzystne. Łatwo to dostrzec,

zwracając uwagę choćby na postępujący proces standaryzacji wielu dziedzin życia. W konsekwencji, nie tylko przestrzeń możliwych sytuacji jest z góry określona, ale również dynamika samego środowiska, w którym realizowane są działania (patrz: szczegółowe procedury postępowania stosowane na lotniskach, np. procedura odprawy biletowo-bagażowej). W tak pojmowanym środowisku obowiązuje szereg deterministycznych reguł, które mają „gwarantować", że uwzględniający je plan zakończy się sukcesem, tzn. doprowadzi do z góry wyznaczonego skutku. Tak funkcjonują procedury w bankach, taśmowa organizacja pracy; tak konstruowane są instrukcje obsługi złożonych urządzeń lub systemów informatycznych. W tego typu zestandaryzowanych środowiskach przewodniki (*guidelines*), o których wspomina Montague, na ogół muszą być bardzo precyzyjne i szczegółowe, gdyż końcowy rezultat często zależy od kolejności, w jakiej zostaną wykonane działania. Nie można, na przykład, wsiąść do samolotu przed nadaniem bagażu, a ten z kolei nie może być nadany, jeśli zawiera w sobie przedmioty zagrażające bezpieczeństwu lotu. We wszystkich tego typu przypadkach charakterystyczna dla środowiska naturalnego niepewność zostaje w znaczący sposób ograniczona – niejako automatycznie podnosząc wartość szczegółowego planowania.

Wiedza domenowa ma również wpływ na treść planu. Na przykład, kiedy pierwszy raz chcemy skorzystać z transportu lotniczego, to na ogół staramy się zgromadzić jak najwięcej informacji o organizacji lotniska, obowiązujących na nim zasadach, o punktach kontroli, itd. Wszystkie niezbędne dane, jak trafnie zauważył Montague, poszerzają nasz mentalny model środowiska. Nie możemy przy ich pomocy zbudować precyzyjnej instrukcji zapewniającej nam osiągnięcie celu, możemy za to skonstruować prowizoryczny scenariusz, który z dużym prawdopodobieństwem będzie przydatny w drodze na lotnisko i przy odprawie. Tak rozumiany plan nie musi być dokładny, wystarczy, że będzie zawierał najważniejsze elementy – pewne węzłowe etapy, których realizacja pomoże w osiągnięciu oczekiwanego rezultatu. Prowizoryczny plan zakłada, że niskopoziomowe zachowania zostaną dynamicznie dodatkowo wygenerowane w trakcie realizacji danego etapu (podcelu). W takich sytuacjach kolejny raz pojawia się potrzeba hierarchizacji zachowań. Aby prowizoryczne plany służyły do osiągania złożonych celów, agent powinien nie tylko dostrzegać związki między sekwencjami poszczególnych działań a osiąganymi stanami świata (np. wiedzieć, że uzyskanie statusu osoby „zakwalifikowanej" do lotu wymaga przejścia przez kontrolę bezpieczeństwa), ale powinien również dysponować reprezentacjami, które umożliwią przeprowadzanie odpowiednich rozumowań. Warto

dodać, że umiejętność planowania wiąże się ze zdolnością do szacowania ryzyka i korzyści/strat związanych z realizacją danego scenariusza. W przypadku, gdy osiągnięcie danego celu obarczone jest wysokim ryzykiem niepowodzenia, wówczas albo rezygnujemy z jego realizacji, albo staramy się przed nim zabezpieczyć, konstruując szczegółowe plany zawierające awaryjne opcje na wypadek, gdyby coś się nie udało[138].

Z powyższych rozważań można wyciągnąć następujący wniosek: obserwacje Reada Montague są niewątpliwie trafne w odniesieniu do środowiska naturalnego, w którym szczegółowe planowanie jest na ogół nieefektywne i kosztowne energetycznie dla mózgu zaangażowanego w realizację niezbędnych procesów poznawczych. W zaawansowanym środowisku cywilizacyjnym planowanie jest niezwykle użytecznym, często najskuteczniejszym narzędziem do efektywnego osiągania celów. Bez niego trudno wyobrazić sobie funkcjonowanie instytucji społecznych, złożonych form organizacji pracy oraz technologii ułatwiających nam codzienne życie.

Racja 3: Plany prowizoryczne są otwarte na zmiany a zarazem stabilne.

Efektywność i prowizoryczność planów wymagają, by tego typu reprezentacje były otwarte na zmiany, by „dostosowywały" się do nowych okoliczności i niespodziewanych stanów środowiska. Każde nowe zdarzenie lub sytuacja, które mogą mieć wpływ na osiągnięcie celu, powinny powodować modyfikację planu lub jego przebudowanie, a w skrajnym przypadku – porzucenie. Kurczowe trzymanie się planu – bez względu na okoliczności – oznaczałoby dodatkowe koszty, a nawet niezdolność do osiągnięcia celu. Jednak otwartość na zmiany nie powinna być nieograniczona. Wspomagany planami układ kontroli zachowań jest skuteczny w osiąganiu celów, jeśli zachowuje stabilność. Brak filtrów służących agentowi do rozróżniania informacji istotnych od nieistotnych mógłby powodować efekty podobne do zbyt częstego przełączania się między celami, tj. do perseweracji lub bardzo wysokiego poziomu rozproszenia (Stuss & Knight, 2002). Trudno określić, jakie zasady decydują o tym, że podsystem planowania dopuszcza wprowadzenie zmian we wcześniej opracowanym planie. Być może, procedura planowania jest po prostu ponawiana, a być może zmiany mają charakter selektywny i dotyczą tylko najbliższych etapów realizacji planu. W każdym razie, w przypadku zmiany planu duże znaczenie mają

[138]Wiele tzw. praktyk zarządczych polega współcześnie na umiejętności oszacowania ryzyka oraz na opracowaniu odpowiednich scenariuszy jego minimalizacji (*PMBOK Guide and Standards | Project Management Institute,* 2018)

odpowiednie heurystyki oraz różnego rodzaju rozumowania, w szczególności te oparte na regule Bayesa.

Dotychczasowe rozważania dotyczące planów oraz ich cech prowadzą do następujących konkluzji. Plany, wraz z prior intencją, umożliwiają agentowi zwiększenie efektywności osiągania celów poprzez usprawnienie metody uczenia ze wzmacnianiem, a dokładniej – poprzez dostarczenie dodatkowych informacji umożliwiających skrócenie czasu eksploracji środowiska, a w konsekwencji szybsze przejście do trybu jego eksploatacji, czyli stosowania optymalnej strategii pozyskiwania nagród. W większości przypadków plany mają charakter wstępny i prowizoryczny, tzn. obejmują jedynie najważniejsze działania związane z realizacją danego celu (wszystkie niskopoziomowe zachowania są dobierane dynamicznie za pomocą podsystemu P-H-RL-OD). Tego typu „podział pracy" podyktowany jest w dużej mierze nieprzewidywalnością oraz złożonością środowiska, nie ma sensu tworzenie detalicznych planów, skoro wiele szczegółów może okazać się niezgodne z naszymi przewidywaniami. Podczas konstrukcji planów, w procesie deliberacji, analizowane są różne scenariusze realizacji celu. Ten, który uznany zostanie za najbardziej korzystny w danym kontekście, poprzez prior intencję będzie włączony w podsystem motywacji i monitorowania. Kiedy nadarzy się okazja, podsystem ten zidentyfikuje odpowiedni cel, uaktywni związany z nim plan i zacznie go realizować. Przebieg procesu dążenia do celu przy wykorzystaniu planu jest zagadnieniem samym w sobie. Wymaga to doprecyzowania relacji między podsystemem planowania i realizacji planów (P-PRP) a podsystemem hierarchicznego uczenia się ze wzmacnianiem z optymalizacją domenową (P-H-RL-OD). Zagadnieniu temu poświęcony zostanie kolejny punkt niniejszego rozdziału.

Doskonalenie mechanizmu kontroli zachowań: od metody prób i błędów do prowizorycznego planowania.

Jeśli porównać proces odpowiedzialny za konstrukcję planów z procesem kontroli zachowań wykorzystującym plany, to ten pierwszy wydaje się być znacznie bardziej złożony. Proces ten wymaga, aby agent posiadał złożoną sieć stanów intencjonalnych, a ponadto – aby opanował on cały szereg zaawansowanych procesów poznawczych, umożliwiających przeszukiwanie poszczególnych fragmentów sieci i budowanie stosownych scenariuszy. Mogłoby się wydawać, że jeśli już agent dysponuje planem, to najtrudniejszy etap konstrukcji działań intencjonalnych został zrealizowany. Wystarczy

tylko, by przekształcił plan w odpowiedni zbiór programów motorycznych, a cały układ zadziała zgodnie z oczekiwaniami (intencją). Wielu badaczy zdaje się nie dostrzegać znaczenia, jakie ma plan (zarówno na etapie jego konstrukcji, jak i na etapie jego realizacji) dla skutecznego wykonania działania intencjonalnego. Patrick Haggard twierdzi na przykład, że aby wyjaśnić działanie intencjonalne, wystarczy wskazać mechanizm transponujący pragnienia i cele na zachowanie[139]. Niestety, poza tym ogólnym stwierdzeniem, nic więcej nie wiadomo na temat działania tego mechanizmu. Nie jest jasne czy tego typu operację należy utożsamić z „dekompresją" informacji zawartych w pragnieniu lub w celu, czy może dałoby się do niej włączyć dyskutowane wyżej elementy planowania polegające na wyznaczeniu kierunków działań, których szczegóły będą „dobierane" w trakcie ich realizacji. Gdyby przyjąć tę drugą interpretację (mechanizm „przekładania" pragnień i celów zawiera także procesy planowania), to zmodyfikowana propozycja Haggarda byłaby zgodna z dookreśleniem przedstawionej wcześniej hipotezy H2:

H2^{d}: Wraz z rozwojem osobniczym kontrola zachowań uzyskuje stopniowo dwupoziomową strukturę: na niższym poziomie dobór zachowań kontrolowany jest przez podsystem hierarchicznego uczenia się ze wzmacnianiem; na wyższym poziomie dobór poszczególnych działań odbywa się na podstawie prowizorycznego planu, który kontrolowany jest przez podsystem realizacji planów.

W H2^{d}, inaczej niż to jest w oryginalnym ujęciu Haggarda, podkreśla się hierarchiczny charakter procesu kontroli zachowań oraz następujące w obrębie tej hierarchii zmiany. Przyjęta konstrukcja uwzględnia dwa składniki: (1) mechanizm uczenia się wbudowany w obydwa podsystemy (P-H-RL-OD oraz P-PRP) oraz (2) stopniowo konstruowany mechanizm rozdzielania kompetencji między wymienionymi dwoma podsystemami. To, iż każdy z tych podsystemów z osobna jest zdolny do konstruowania nowych reprezentacji (np. zachowań wyższego rzędu, przekonań na temat praw rządzących środowiskiem, w którym realizowane są działania), umożliwia ich rozwój. Odrębność tych podsystemów sprawia, że każdy z nich inaczej wpływa na dobór zachowań oraz na ich kontrolę. Dlatego też, zanim scharakteryzuję związki między nimi, opiszę etapy rozwoju każdego z nich z osobna.

[139] „Termin «intencja» odnosi się do kilku różnych procesów w obrębie łańcucha przetwarzania informacji, przekładających pragnienia i cele na zachowanie." (Haggard, 2005, s. 290).

Porządek prezentacji przedstawia się następująco: najpierw określone zostaną główne fazy rozwojowe podsystemu hierarchicznego uczenia się ze wzmacnianiem z optymalizacją domenową (P-H-RL-OD), następnie główne fazy rozwojowe podsystemu planowania i realizacji planów (P-PRP), a na końcu naszkicowana zostanie dynamika współpracy między tymi podsystemami. Opis każdego z podsystemów zawierać będzie: stan początkowy, rozwój oraz stan dojrzały.

Główne fazy rozwojowe podsystemu hierarchicznego uczenia się ze wzmacnianiem z optymalizacją domenową (P-H-RL-OD)

Stan początkowy

W proponowanym modelu 3.0 zakłada się, że agent od urodzenia dysponuje pewnym bazowym układem mechanizmów umożliwiających wykorzystanie wrodzonych reprezentacji niezbędnych do działania mechanizmu uczenia się ze wzmacnianiem. Tego rodzaju układ obejmuje:

- o – konstruowanie obserwacji, na podstawie których możliwe jest wyznaczenie stanów świata ‘s’,
- z – zdolność do reprezentowania elementarnych zachowań,
- $r_{ś}$ / r_h – zdolność do reprezentowania nagród uzyskiwanych ze środowiska,
- V – zdolność do konstrukcji funkcji wartości,
- π – zdolność do stosowania określonej strategii doboru zachowań.

Powyższa lista reprezentacji pozwala agentowi zainicjować proces realizacji celu. Na tym etapie cel jest określonym typem nagrody, której pozyskiwanie aktywuje podsystem monitorowania i motywacji (P-MM) reagujący m.in. na zaburzony stan homeostazy oraz na określone bodźce pochodzące z otoczenia. Przykładem tego typu celu może być pragnienie zaspokojenia głodu, chęć napicia się wody, potrzeba zabawy, itd. Zgodnie z hipotezą bramkowania dopaminergicznego podsystem (P-MM) filtruje napływające obserwacje, a co za tym idzie – stabilizuje proces realizacji celu. Z jednej strony, realizowany aktualnie cel powinien mieć zagwarantowany odpowiednio długi czas realizacji, z drugiej strony, powinien być „wrażliwy” na zmieniające się okoliczności, w szczególności na niebezpieczeństwa lub okazje.

W związku z powyższym, stan początkowy podsystemu (P-H-RL-OD) można przedstawić jako następujący układ reprezentacji: (o, z, $r_{ś}$ / r_h, s, V, π).

Rozwój

Sukcesy i porażki w pozyskiwaniu podstawowych typów nagród, niezbędnych do utrzymania homeostazy oraz zaspokojenia wrodzonych potrzeb organizmu (np. eksploracji), prowadzą do poprawy efektywności realizacji celów, są również źródłem nowych reprezentacji, tzw. zachowań wysokiego poziomu (Z) oraz uogólnionych obserwacji (O). To z ich pomocą podsystem (P-H-RL-OD) może „przenosić" nabyte umiejętności takie jak: posługiwanie się narzędziami, omijanie przeszkód, sterowanie urządzeniami technicznymi, itp., z jednego celu na inny. Wielopoziomowa struktura zachowań ułatwia również optymalizację tych umiejętności. Poprawa danej umiejętności w warunkach X prowadzi do poprawy w warunkach Y. Ponadto, tego typu reprezentacje uzyskują – wraz z rozwojem sieci stanów intencjonalnych – swoją specyficzną treść oraz nakierowanie na zgodność, innymi słowy, uzyskują swój intencjonalny wymiar i stają się „pełnoprawnymi" „członkami" sieci. W ten sposób nie tylko „organizują" one sekwencje zachowań elementarnych, ale zyskują również możliwość uczestniczenia w procesach deliberacyjnych (patrz schemat przebiegu prostego działania intencjonalnego wg Searle'a zaprezentowany został w rozdziale 3.), które posłużą w stanie dojrzałym do wyznaczania planów. Na tym etapie można również spodziewać się zachowań odpowiedzialnych za pozyskiwanie nagród (R_{SI}), których bazą są odpowiednio zwartościowane stany intencjonalne, np. pragnienia (patrz: hipoteza „super power" Reada Montague).

Rozwój podsystemu (P-H-RL-OD) polega na tym, że jego stan początkowy wzbogacony zostaje o reprezentacje typu O, Z oraz R_{SI}. W całości tak rozwinięty układ reprezentacji można przedstawić w następujący sposób: (o, **O**, z, **Z**, $r_{ś}$ / r_h, **R_{SI}**, s, V, π).

Stan dojrzały

W pełni rozwinięty podsystem (P-H-RL-OD) potrafi efektywnie stosować zhierarchizowany układ zachowań, a także otwarty jest na wiedzę domenową zawartą w sieci stanów intencjonalnych. Wiedza ta nie jest bezpośrednio dostępna dla algorytmów RL i, zanim zostanie wykorzystana, jest „sprowadzona" do postaci tzw. nagród kształtujących (*shaping rewards*). Dopiero wtedy podsystem uczenia się ze wzmacnianiem może z niej skorzystać i w efekcie skrócić czas fazy eksploracji, czyli okres, w którym agent aktywnie uczy się poruszania w środowisku oraz poszukuje optymalnej strategii doboru zachowań. Przyjęte rozwiązanie pozwala zachować kluczowe własności algorytmu uczenia się ze wzmacnianiem (m.in. zdolność do adaptacji, umiejętność „odkrycia"

optymalnej strategii doboru zachowań, itp.) i równocześnie połączyć go z podsystemem planowania, który – wykorzystując wiedzę domenową zgromadzoną w (P-ZSSI) może „kształtować" i oddziaływać na podsystem doboru zachowań elementarnych. Słowo „kształtuje" dobrze oddaje relację między podsystemem planowania a podsystemem uczenia się ze wzmacnianiem, nie jest to bowiem prosty związek przyczynowo-skutkowy, ale złożona relacja międzysystemowa. Warto dodać, że informacja dostarczana za pomocą nagród kształtujących nie jest jedynym czynnikiem, który wpływa na funkcję wartości, a co za tym idzie na wybór zachowań elementarnych. Na wybór takich zachowań, oprócz nagród kształtujących, które są wyznaczone na etapie planowania, wpływ mają również: funkcja wartości reprezentująca wcześniejsze doświadczenia (charakterystyczna dla algorytmu uczenia się ze wzmacnianiem), współczynnik dyskonta[140], a także wartość nagród pozyskiwanych ze środowiska (patrz: opis algorytmu TDRL zaprezentowany w rozdziale 3.).

Stany intencjonalne, które uzyskały status nagrody, są drugim ważnym „interfejsem" łączącym podsystem planowania z podsystemem uczenia się ze wzmacnianiem. W prezentowanym wyżej modelu 1.2 – w nawiązaniu do hipotezy „nad-mocy" (*super power*) Montague – przyjęto, że odpowiednio rozszerzone pojęcie nagrody pozwala wyjaśnić obserwowane u zwierząt nowe formy zachowań. Przypisanie aktowi behawioralnemu (np. tańcowi godowemu) wysokiej wartości przez układ nerwowy „gwarantuje", że tego typu akt będzie powtarzany w odpowiednich kontekstach. Podobnie, jak twierdzi Montague, mogą funkcjonować inne typy reprezentacji, także te bardziej abstrakcyjne, na przykład chęć zdobycia prestiżowej nagrody w matematyce czy nadzieja na przeniesienie się do innego wymiaru rzeczywistości wzbudzona w członkach sekty *Heaven's Gate* przez ich założycieli. Rozciągnięcie wskazanej dyspozycji na niemal dowolny typ pragnienia, które jest dobrze osadzone w sieci stanów intencjonalnych, powoduje, że repertuar naszych zachowań wzrasta niepomiernie, staje się zdecydowanie bogatszy, niż u zwierząt i umożliwia niemal nieograniczoną różnorodność celów.

Nietrudno zauważyć, porównując podsystemy uczenia się ze wzmacnianiem przedstawione w modelach 1.2 i 3.0, że główna zmiana polegała na dodaniu w modelu 3.0 funkcji umożliwiającej wykorzystanie wiedzy domenowej do optymalizacji zachowań (stąd przyrostek OD – optymalizacja domenowa). Tego typu rozszerzenie pozwala

[140] γ to współczynnik dyskonta ($\gamma \leq 1$) powodujący, że ta sama nagroda otrzymywana z opóźnieniem jest dla agenta mniej wartościowa, niż nagroda otrzymana wcześniej (P. Cichosz, 2007, s. 718).

utrzymać autonomię podsystemu (P-H-RL-OD) – przy równoczesnej otwartości na wiedzę zgromadzoną w innej postaci i w innym podsystemie.

Dojrzała wersja podsystemu (P-H-RL-OD) uzyskuje ostatecznie postać, którą reprezentuje następujący układ: (o, O, z, Z, $r_{ś}$ / r_h, R_{SI}, $\mathbf{r_k}$, s, V, π).

Przedstawione powyżej fazy rozwoju podsystemu (P-H-RL-OD) można przestawić symbolicznie w następujący sposób:

(o, z, $r_{ś}$ / r_h, s, V, π) → (o, **O**, z, **Z**, $r_{ś}$ / r_h, $\mathbf{R_{SI}}$, s, V, π) → (o, O, z, Z, , $r_{ś}$ / r_h, , R_{SI}, r_k, s, V, π, SI)

Powyższa sekwencja wyraźnie wskazuje, jak pojawianie się nowych typów reprezentacji oraz powiązanych z nimi mechanizmów ich konstruowania, przetwarzania i stosowania prowadzi do coraz bardziej złożonych form działania. Przy pomocy tego zestawienia można również rozpoznać, co w przybliżeniu może się stać, gdy dojdzie do zaburzenia rozwoju zdolności poznawczych (patrz: przypadek dzieci wychowywanych w izolacji wskazany w kontekście modelu 2.0).

Główne fazy rozwojowe podsystemu planowania i realizacji planów (P-PRP)

Stan początkowy

Podsystem planowania i realizacji planów początkowo działa na ograniczonych zasobach. Przez wiele miesięcy, a być może nawet lat, trwa proces uczenia się. Ten stosunkowo długi okres w głównej mierze jest konsekwencją zależności, która występuje między planowaniem a siecią stanów intencjonalnych. Trawestując słynny slogan: *no computation without representation*, można by powiedzieć: bez reprezentacji nie ma planowania. Innymi słowy, dopóki agent nie nabędzie zdolności do tworzenia odpowiednio złożonych reprezentacji oraz rozmaitych form manipulowania nimi, dopóty nie będzie mógł wykorzystywać tego typu wiedzy do organizacji działań. Znaczy to, że podsystem (P-PRP) przez stosunkowo długi okres funkcjonuje w trybie „eksploracji" a nie eksploatacji, jak określiliby to badacze zajmujący się metodą uczenia się ze wzmacnianiem. Innymi słowy, w stanie początkowym podsystem planowania wyłącznie „wytwarza" reguły określające jak wykorzystywać reprezentacje zgromadzone w sieci stanów intencjonalnych do organizacji działań, ale ich nie stosuje do kontroli zachowań. W tym czasie rzeczywista

kontrola zachowań odbywa się za pośrednictwem mechanizmu uczenia się ze wzmacnianiem, który, jak to zostało wcześniej powiedziane, dostępny jest od urodzenia.

Do tej pory nie udało się ustalić, jak w szczegółach przebiega proces wykształcania się reguł niezbędnych do konstrukcji planu w podsystemie (P-PRP) i co decyduje, że w pewnym momencie rezultaty jego działania, czyli specyficzne reprezentacje, zostają włączone w mechanizm uczenia się ze wzmacnianiem (patrz: koncepcje tzw. opcji oraz nagród kształtujących opisane w rozdziale 3.). Możemy tylko spekulować, że reguły te mają związek z postulowaną przez Searle'a przyczynowością intencjonalną oraz zbliżoną do niej na poziomie koncepcyjnym, choć nie ontologicznym, pozorną mentalną przyczynowością Daniela Wegnera, ufundowaną na zasadzie priorytetu, spójności i wyłączności. Wskazane koncepcje w podobny sposób tłumaczą, jak agent interpretuje związek pomiędzy oczekiwanym efektem działania, podjętym zachowaniem a uzyskanymi rezultatami. Przywołane przez Wegnera dane eksperymentalne zdają się potwierdzać, że przeżycie sprawstwa to wytwór nieświadomego wnioskowania uwzględniającego wymiar czasowy zdarzeń powiązanych z działaniem (zasada priorytetu), ich związek z wiedzą podmiotu (zasada spójności) oraz kontekstem społecznym (zasada wyłączności). Złożona natura tego wnioskowania sugeruje jednak, że nie jest ono wrodzone. Trudno bowiem przypuszczać, by przekonania należące do wiedzy potocznej z tzw. psychologii czy fizyki ludowej, zakładane przez model Wegnera, były nam dostępne od pierwszych dni życia. Ponadto, procesy niezbędne do przeprowadzenia wskazanego wnioskowania są na tyle złożone, że wymagają odpowiednio zaawansowanych funkcji wykonawczych (w szczególności pamięci roboczej), które same kształtują się po odpowiednie długim treningu. W tej sytuacji należy uznać, że wnioskowanie implikowane przez hipotezę Wegnera zaczyna sprawnie funkcjonować w obrębie podsystemu zarządzania siecią stanów intencjonalnych (P-ZSSI) oraz planowania i realizacji planów (P-PRP) dopiero wtedy, kiedy sieć jest już odpowiednio złożona, a zasoby poznawcze niezbędne do manipulowania jej składnikami są wystarczająco rozwinięte.

Podpowiedzią, jak (P-PRP) oraz (P-ZSSI) mogą funkcjonować na wczesnym etapie rozwoju, jest idea wykorzystana w głębokim uczeniu się ze wzmacnianiem (*deep reinforcement learning*), czyli w jednej z najnowszych implementacji algorytmu RL[141].

[141] Najbardziej spektakularnym sukcesem głębokiego uczenia się ze wzmacnianiem jest zwycięstwo systemu Alpha-GO nad mistrzem świata Lee Sedolem (Silver i in., 2016).

Powtarzanie doświadczenia[142] (*experience replay*), bo tak nazywa się ta idea, polega na wykorzystaniu sztucznej sieci neuronowej do modelowania funkcji, która na podstawie zadanego stanu świata (s) wyznaczy wartość zachowania dla tego stanu: Q(s, z)[143]. Zgodnie z zasadą działania algorytmu uczenia się ze wzmacnianiem (RL) funkcja Q(s,z) pozwala agentowi realizować optymalną strategię doboru zachowań (wynika to wprost z tzw. zasady optymalności Bellmana). Co szczególnie istotne, doświadczenia zapamiętywane podczas eksploracji środowiska zostają poprzez wykorzystanie sieci neuronowej **uogólnione**, tzn. przybliżana przez sieć neuronową funkcja Q umożliwia agentowi prawidłowo szacować wartość zachowań również dla nieznanych mu stanów środowiska.

Reprezentacje wykorzystywane w procedurze powtórnego doświadczania (Zhao i in., 2016) mają następującą strukturę: (s_t, z_t, s_{t+1}, r_{t+1}). Nietrudno zauważyć, że ich typy oraz układ są zbieżne z zaproponowanym w zintegrowanym modelu działań intencjonalnych układem reprezentacji przesyłanych z podsystemu uczenia się ze wzmacnianiem (P-H-RL-OD) do podsystemu zarządzania siecią stanów intencjonalnych (P-ZSSI) - : s_t, z_t, $Z_t \rightarrow s_{t+1}$, δ_{t+1}, R_{SI}, , $r_ś$ / $r_{h\,t+1}$, o_{t+1}, O_{t+1}. Wskazany układ reprezentacji jest nieco bogatszy, niż ten, który wykorzystany został w algorytmie głębokiego uczenia się ze wzmacnianiem (*deep reinforcement learning*), gdyż obejmuje on również reprezentacje wysokiego poziomu, takie jak: Z, R_{SI}, O oraz błąd predykcji nagrody δ. Jednak co do istoty, obydwa podejścia operują podobnym zakresem informacji i pełnią podobną funkcję, tzn. służą do poszukiwania regularności funkcjonujących w środowisku, których znajomość umożliwia przewidywanie skutków zachowań, a docelowo również konstruowanie planów. Równocześnie należy stwierdzić, że mechanizm gromadzenia doświadczeń oraz wyznaczania na ich podstawie funkcji Q to tylko inspiracja, gdyż oferowana przez ten mechanizm generalizacja doświadczeń jest z perspektywy działań intencjonalnych zbyt wąska. Przypomnę, że złożone działania intencjonalne bazują na sieci stanów intencjonalnych, a ta, jak sama nazwa wskazuje, wiąże bardzo różne doświadczenia, nie

[142] Można przypuszczać, że „powtórzone doświadczenie" (*experience replay*) to bezpośrednie nawiązanie do badań Matthew Wilsona, który przy pomocy metod neuronalnego dekodowania odkrył, że szczury podczas snu odtwarzają w tzw. komórkach miejsca (*place cells*) niedawno nabyte doświadczenia z labiryntu (T. J. Davidson i in., 2009).

[143] Należy dodać, że sieć neuronowa wykorzystana do modelowania funkcji Q to tzw. rozwiązanie przybliżone. Innymi słowy, w określonych przypadkach funkcja Q może dla danego stanu błędnie szacować wartość danego zachowania.

tylko te, które pozyskane zostały w ramach pojedynczego celu, jak to ma miejsce w głębokim uczeniu się ze wzmacnianiem[144].

Podsumowując, w zintegrowanym modelu działań intencjonalnych zakłada się, że w stanie początkowym podsystem planowania i realizacji planów nie posiada żadnych specyficznych typów reprezentacji. Podsystem ten dysponuje jedynie bazowymi regułami organizującymi doświadczenia gromadzone w podsystemie zarządzania siecią stanów intencjonalnych (P-ZSSI). Na tym etapie rozwoju podsystemu (P-PRP) zakłada się, że rezultaty jego działania **nie są** wykorzystywane do kontroli zachowań.

Rozwój

Proponuję przyjąć, że podmiot uzyskuje zdolność planowania w momencie, gdy potrafi świadomie utworzyć najprostszą prior intencję, a następnie zrealizować działanie intencjonalne wynikające z jej warunków spełniania. Zgodnie ze schematem Searle'a prior intencja to rezultat działania procesu deliberacji, czyli procesu poznawczego działającego w odpowiednio złożonej sieci stanów intencjonalnych (np. wnioskowanie, którego przesłankami są wcześniejsze doświadczenia, a konkluzją planowane działanie oraz jego przewidywany rezultat). Jak już zasygnalizowano, zanim podsystem (P-PRP) zacznie „współkontrolować" zachowania podmiotu, musi przejść odpowiedni „trening", dysponować określonymi zasobami (funkcjami wykonawczymi) oraz bazą wiedzy w postaci sieci stanów intencjonalnych (m.in. przekonań należących do fizyki i psychologii ludowej). Możemy tylko spekulować (patrz: koncepcja powtarzania doświadczeń), jak przebiega tego typu trening. To, co oferuje zintegrowany model działań intencjonalnych, to określenie minimalnego zbioru reprezentacji niezbędnego do utworzenia najprostszej prior intencji. By tego typu stan intencjonalny mógł zaistnieć, wymagane są dwie składowe: (1) predykcyjna oraz (2) motywacyjna. Pierwsza określona jest przez treść typowego zamiaru. Jest to spodziewany stan świata po wykonaniu określonego zachowania

[144] Warto w tym miejscu jeszcze raz wspomnieć o aktywnym wnioskowaniu (*active inference*), tj. o propozycji Karla Fristona oraz współpracowników. W podejściu tym próbuje się uogólnić metodę uczenia się ze wzmacnianiem i rozciągnąć ją na obszar procesów poznawczych przy pomocy formalizmu modeli generatywno-rozpoznawczych (patrz: idea mózgu bayesowskiego) oraz przy pomocy zasady minimalizacji swobodnej energii (*free energy principle*) (Kaplan & Friston, 2018). Inną niezwykle intrygującą propozycją jest opublikowany w 2021 roku przez Davida Silvera, Satindera Singha, Doina Precup i Richarda S. Suttona artykuł zatytułowany *Reward is enough*. Wymienieni badacze stawiają w nim następującą tezę: „In this article we hypothesise that intelligence, and its associated abilities, can be understood as subserving the maximisation of reward. Accordingly, reward is enough to drive behaviour that exhibits abilities studied in natural and artificial intelligence, including knowledge, learning, perception, social intelligence, language, generalisation and imitation." (Silver i in., 2021).

lub sekwencji zachowań. Na poziomie reprezentacji tego typu składową można w najprostszej postaci wyrazić jako asocjację (Z_t, s_{t+1}), gdzie Z_t to zachowanie wysokiego poziomu w chwili t, a s_{t+1} to stan świata będący skutkiem zachowania Z_t. Z kolei druga składowa jest niezbędna, by przewidywany stan świata miał dla agenta wartość, „by był sens się trudzić". Stąd niezbędna jest druga asocjacja (Z_t, R_{t+1}), która po „zainstalowaniu" w podsystemie monitorowania prowadzi, po jej aktywowaniu, do zainicjowania odpowiedniego celu w podsystemie (P-H-RL-OD). Warto podkreślić, że tego typu planowanie na poziomie poznawczym nie wymaga złożonych form manipulowania przekonaniami, wystarczą wskazane asocjacje, które na poziomie świadomości wyrażone zostaną w formie zamiaru, którego treść będzie w przybliżeniu następująca: „w chwili t wykonam zachowanie Z".

Główne *novum* związane z wykorzystaniem prior intencji polega na czasowym „odłączeniu" przewidywanych skutków działania od jego realizacji. Tego typu separacja nie jest, niestety, dostępna w metodzie uczenia się ze wzmacnianiem, dlatego agent w bardzo ograniczonym zakresie może decydować o priorytecie poszczególnych celów. W pewnym sensie o globalnym rozkładzie celów decydują określone rytmy związane z homeostazą oraz bodźce napływające ze środowiska. Zdolność do projektowania działań „na przyszłość" przynosi podmiotowi szereg korzyści (patrz: przedstawione powyżej racje uzasadniające przydatność prowizorycznego planowania), równocześnie jej realizacja oparta na współpracy kilku podsystemów powoduje co najmniej dwa problemy: (1) problem synchronizacji oraz (2) problem konkurowania o zasoby. Przykładowo, by proces deliberacji mógł być skutecznie zrealizowany, podsystem uczenia się ze wzmacnianiem powinien na jakiś czas zawiesić swoje działanie lub je znacząco ograniczyć. Najbardziej predestynowany do „podejmowania" tego typu decyzji jest podsystem monitorowania i motywacji (P-MM) odpowiedzialny za aktywację i dezaktywację celów oraz filtrowanie napływających informacji (patrz: hipoteza bramkowania dopaminergicznego opisana w rozdziale 3.). Choć deliberacja to czynność umysłowa, a nie sekwencja zachowań, to z perspektywy całego organizmu obydwa typy aktywności nie powinny funkcjonować niezależnie, gdyż może to prowadzić do niekorzystnych interakcji, np. zamyślenie lub rozmowa na ogół prowadzi do spadku skupienia uwagi podczas jazdy samochodem, co w określonych warunkach może prowadzić do niebezpiecznych sytuacji. Wyraźnie pokazują to badania nad wielozadaniowością (*multitasking*), które podważają mit podzielnej uwagi, wskazując, że jedynie zadania rutynowe mogą być do pewnego stopnia realizowane

równolegle z zadaniami wymagającymi wysiłku poznawczego (Christine Rosen, 2008). W związku z tym zakłada się, że w zintegrowanym modelu działań intencjonalnych to podsystem (P-MM) decyduje, która z czynności, w jakim zakresie oraz z jakimi zasobami będzie realizowana.

O ile w przypadku stanu początkowego podsystemu planowania i realizacji planów zintegrowany model zakłada, że podsystem ten nie posiada żadnych specyficznych reprezentacji, o tyle w fazie rozwoju można przyjąć, że funkcjonująca w obrębie tego podsystemu reguła konstrukcji prior intencji opiera się na co najmniej dwóch typach asocjacji przechowywanych w podsystemie (P-ZSSI): (Z_t, s_{t+1}) oraz (Z_t, R_{t+1}). Znaczy to, że realizacja działania intencjonalnego wymaga współpracy co najmniej trzech podsystemów: (P-PRP), (P-ZSSI), (P-H-RL-OD). Za ich czasową koordynację w zintegrowanym modelu działań intencjonalnych odpowiada podsystem monitorowania i motywacji (P-MM).

Stan dojrzały

Przedstawiony powyżej rozwój podsystemu (P-PRP) w ograniczony sposób wpływa na efektywność doboru działań. Związany z tym stanem minimalny układ reprezentacji nie pozwala efektywnie planować realizacji celów w dłuższym okresie, dlatego w świecie zwierząt trudno dostrzec jego przejawy lub odróżnić działania zaplanowane od zachowań bazujących na mechanizmie uczenia się ze wzmacnianiem, działającym na podstawie dostępnych tu i teraz informacji (obserwacji (o/O) i nagród (r/R)) oraz doświadczeń zakodowanych w formie funkcji wartości. Podobnie, jak się wydaje, funkcjonują dzieci, które w pierwszych latach życia niemal całkowicie skupione są na chwili obecnej, często nie są w stanie przewidzieć skutków swoich działań, a tym bardziej ich zaplanować (Wegner, 2002, s. 22). Wyraźna zmiana następuje w momencie, kiedy dziecko dysponuje odpowiednio rozbudowaną siecią stanów intencjonalnych i zaczyna posługiwać się językiem oraz dysponuje odpowiednio rozwiniętą teorią umysłu (Reuter, 2014). Warto nadmienić, że za rozwój sieci stanów intencjonalnych w przyjętym modelu odpowiada w dużej mierze „transfer" reprezentacji z podsystemu (P-H-RL-OD) do podsystemu (P-ZSSI) (s_t, z_t, $Z_t \rightarrow s_{t+1}$, δ_{t+1}, R_{SI}, $r_ś$ / $r_{h\,t+1}$, o_{t+1}, O_{t+1}), które są aktywowane w trakcie interakcji agenta ze środowiskiem, umożliwiają podsystemowi zarządzającemu siecią stanów intencjonalnych tworzenie przekonań, pragnień i innych stanów intencjonalnych. Z przytoczonego wcześniej cytatu z Łurii dowiadujemy się, że język to nie tylko narzędzie

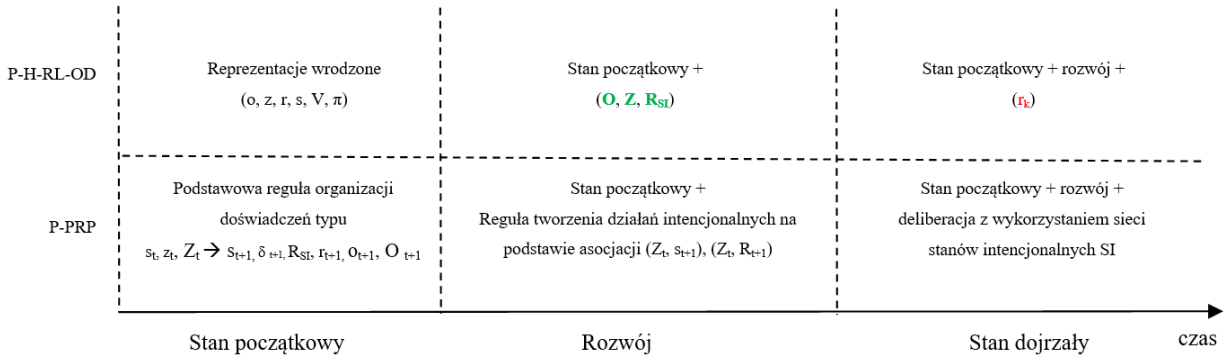
komunikacji lub działania[145], ale również „mechanizm" organizowania zachowań w celowe sekwencje. Kompozycyjność języka umożliwia kreowanie zarówno złożonych wypowiedzi, jak i alternatywnych scenariuszy, w których zakłada się realizację celu. Plan skonstruowany przy wykorzystaniu narzędzia, jakim jest język, stanowi dobrze wyodrębnioną jednostkę, którą łatwo jest poddać ocenie i oszacować prawdopodobieństwo sukcesu jego realizacji. Trudno obecnie określić, które typy operacji mentalnych są w głównej mierze zaangażowane w konstrukcję planu oraz jego ocenę, należy jednak założyć, zgodnie z analizą Montague, że plany nigdy nie zyskałyby tak dużego wpływu na nasze zachowania, gdyby nie ich efektywność w odniesieniu do mechanizmu uczenia się ze wzmacnianiem. Nie zmienia to faktu, że koszt tworzenia planu musi być odpowiednio skalkulowany i zestawiony ze spodziewanymi zyskami z jego realizacji. Zbyt długi okres planowania nie tylko jest kosztowny (bo wymaga wydatkowania znacznej ilości energii w związku z pracą mózgu), ale również znacznie ogranicza zaspokajanie bieżących potrzeb. Dlatego też decyzja o przystąpieniu do przygotowywania planu, jak i wybór poziomu jego szczegółowości jest niezależnym problemem decyzyjnym, którego rozwiązanie wymaga określonego procesu obliczeniowego z funkcją uczenia się i optymalizacji. Ostatecznie zatem, w „dojrzałej" postaci podsystem planowania i realizacji planów, wykorzystując dostępną sieć stanów intencjonalnych, pozwala agentowi na tworzenie mniej lub bardziej szczegółowych planów, instalowanie ich w formie prior intencji w podsystemie motywacji i monitorowania, a następnie – w sprzyjających okolicznościach – na ich aktywację i realizację.

Następujące elementy służą odróżnieniu dojrzałej formy planowania od jej najprostszej postaci:

- Dostępność rozbudowanej sieci stanów intencjonalnych, znaczniej bardziej złożonej, aniżeli proste asocjacje typu (Z_t, s_{t+1}) oraz (Z_t, R_{t+1}).
- Złożoność procesów deliberacyjnych opartych w dużym stopniu na językowym dostępie do sieci przekonań (tzw. wiedzy domenowej).
- Zaawansowane procesy decyzyjne i heurystyki służące określeniu zakresu planowania oraz jego szczegółowości.

[145] Język również może pełnić, zgodnie z tzw. koncepcją aktów mowy Austina, funkcje performatywne (Green, 2017).

Przedstawioną powyżej analizę można zobrazować przy pomocy następującego schematu:



Rys. 6. Rozwój podsystemu hierarchicznego uczenia się ze wzmacnianiem z optymalizacją domenową (P-H-RL-OD) oraz podsystemu planowania i realizacji planów (P-PRP).

Należy zaznaczyć, że powyższy schemat ma charakter poglądowy. Główne uproszczenie polega na tym, że stany obu podsystemów następują w tym samym czasie, w rzeczywistości przejścia między stanami mogą się odbywać w różnych momentach i do pewnego stopnia niezależnie od siebie.

Dynamika związków między podsystemem planowania (P-PRP) a podsystemem uczenia się ze wzmacnianiem (P-H-RL-OD).

Rola podsystemów P-PRP oraz P-H-RL-OD zmienia się w czasie. Najpierw dominuje podsystem uczenia się ze wzmacnianiem, który w początkowych fazach rozwoju – przy pomocy ograniczonych zasobów „reprezentacyjnych" – realizuje cztery ważne zadania: (1) kontroluje zachowania niezbędne do zaspokojenia podstawowych potrzeb, (2) przekształca wrodzone odruchy w zachowania wyższego poziomu, (3) konstruuje dyspozycje tła, (4) na wejście podsystemu zarządzania siecią stanów intencjonalnych „dostarcza" elementarne warunki spełniania, przy pomocy których możliwa staje się konstrukcja pragnień, przekonań, obaw, lęków, itp. Równolegle rozwija się i funkcjonuje podsystem planowania, który początkowo wspiera podsystem uczenia się ze wzmacnianiem, a z czasem zaczyna go w znacznej mierze kontrolować. W początkowych fazach planowanie sprowadza się do prostych sekwencji zachowań wysokopoziomowych lub do wyznaczania sekwencji nagród niezbędnych do realizacji celu. Z czasem planowanie rozszerza się i przybiera postać prowizorycznych scenariuszy opartych na wiedzy dziedzinowej, np. by dotrzeć do szkoły, najpierw muszę dojść do przystanku autobusowego, następnie wsiąść do autobusu nr 5,

skasować bilet, wysiąść na odpowiednim przystanku, itd. Tak opracowane plany poddawane są ewaluacji. Najbardziej efektywny z nich – w zależności od okoliczności – jest „instalowany" w podsystemie (P-MM) i aktywowany w odpowiednim momencie. Na tym etapie można mówić już o wyraźnej dominacji planowania nad mechanizmem uczenia się ze wzmacnianiem. Nadal jednak każdy z podsystemów zachowuje swoją autonomię oraz logikę działania. Na tym etapie ich współpraca opiera się na następującej pętli zwrotnej:

- (P-PRP) „wysterowuje" (P-H-RL-OD) w taki sposób, by maksymalnie skrócić czas eksploracji,
- (P-H-RL-OD) za pośrednictwem (P-ZSSI) „dostarcza" informacje do (P-PRP), są one niezbędne w realizacji planu lub w jego modyfikacji.

W ten sposób cały układ złożony ze wskazanych podsystemów zyskuje możliwość nabywania nowych kompetencji i realizacji coraz bardziej złożonych celów, w tym dalekosiężnych, jak np. planowanie kariery zawodowej czy tworzenie strategii rozwoju firm.

Podsumowanie

Przedstawiona w niniejszym punkcie analiza procesu doskonalenia mechanizmu kontroli zachowań wieńczy rozważania dotyczące zintegrowanego modelu działań intencjonalnych. Wprowadzony do modelu 3.0 podsystem planowania i realizacji planów w zasadniczy sposób zmienia organizację działań intencjonalnych. W wymiarze czasowym cel zostaje podzielony na fazę przygotowawczą oraz fazę realizacji. W praktyce oznacza to, że podmiot może wybierać moment oraz stan środowiska, w którym pozyskanie nagród (realizacja celu) będzie – z jego perspektywy – najłatwiejsze albo wartość celu najkorzystniejsza. Cechy tej nie posiada metoda uczenia się ze wzmacnianiem, stąd jej reaktywny charakter.

Kolejną ważną cechą nowego trybu kontroli zachowań jest jego związek z wiedzą domenową, która w zasadniczy sposób zmienia relację agent – środowisko. Nawet niedoskonałe i uproszczone *know-how* z zakresu psychologii czy fizyki ludowej pozwala wykorzystać określone związki przyczynowe obecne w świecie do realizacji własnych zamiarów. Posługiwanie się narzędziami to typowy przykład tego typu „manipulacji" w

środowisku. Co istotne, zintegrowany model działań intencjonalnych zakłada, że planowanie oraz powiązana z nim sieć stanów intencjonalnych nie stanowią zupełnie odrębnego i niezależnego od podsystemu (P-H-RL-OD) trybu kontroli zachowań. Przeciwnie, w modelu tym przyjmuje się, że podsystem planowania i realizacji planów uczy się, jak tego typu mechanizm wykorzystać i rozszerzyć. Koncepcja zachowań wysokiego poziomu (tzw. opcji - Z) oraz nagród kształtujących (r_k) to dwa kluczowe typy reprezentacji umożliwiające współpracę między nimi. Przyjęte rozwiązanie ma dwie zalety. Po pierwsze, tak zorganizowaną kontrolę zachowań cechuje swoista „ciągłość", to znaczy, że mechanizm kontroli wykorzystywany na wczesnych etapach rozwoju jest również stosowany w fazie dojrzałej. Innymi słowy, planowanie nie zastępuje kontroli zachowań zależnej od metody uczenia się ze wzmacnianiem, ale ją wspomaga i udoskonala. Drugą zaletą zaproponowanego w zintegrowanym modelu podejścia jest hierarchiczny „podział pracy" między podsystemami: (P-PRP) oraz (P-H-RL-OD). Można powiedzieć, że w tym układzie (P-PRP) skupia się na kwestiach „strategicznych" (patrz: koncepcja planów prowizorycznych), a (P-H-RL-OD) na szczegółach, które trzeba „wygenerować" w trakcie realizacji planu. To, że podsystem hierarchicznego uczenia się ze wzmacnianiem działa w trybie ciągłej oceny uzyskiwanych rezultatów (patrz: hipoteza dopaminergicznego błędu predykcji nagrody), sprawia, że podsystem planowania jest niejako „zabezpieczony" przed projektowaniem działań wysoce nieefektywnych, trudno bowiem sobie wyobrazić, by w normalnych warunkach podsystem planowania ignorował przesyłane przez podsystem (P-H-RL-OD) błędy predykcji nagrody (wyjątkiem jest w tym przypadku stan uzależnienia, patrz: analiza zawarta w rozdziale 3.).

6 Zakończenie

Bret Weinstein, biolog ewolucyjny, twierdzi:

> *Z jednej strony, człowiek, jak każdy wytwór ewolucji, jest na swój sposób wyjątkowy, jednak «wyjątkowość» naszego gatunku jest szczególna, polega mianowicie na tym, że znacząco większy procent repertuaru naszych zachowań – w porównaniu z innymi gatunkami – został przeniesiony na warstwę kulturową. To wzajemne oddziaływanie między naszymi genomami, które pod wieloma względami są dość standardowe, i warstwą kulturową nie występuje u żadnych innych stworzeń na Ziemi. Tego typu oddziaływanie dla kogoś, kto myśli ewolucyjnie, jest ważne i musi być traktowane z należytą uwagą, równocześnie jest ono kluczowe dla naszego istnienia, gdyż umożliwia nam coś, czego żadne inne stworzenie na Ziemi nigdy nie było w stanie zrobić.*[146] (Cytat z wypowiedzi wygłoszonej podczas konferencji Virtual Futures, 2018).

Powyższa uwaga dobitnie pokazuje, iż – zdaniem cytowanego autora – niezwykle trudno oddzielić wpływ czynników biologicznych od wpływu czynników kulturowych na ludzkie zachowania. W tej, z pozoru oczywistej, konstatacji pomija się kwestię zasadniczą. To mianowicie, że warstwa kulturowa ma charakter społeczny, czyli ponadjednostkowy. Znaczy to, że czynniki kulturowe mogą – w odróżnieniu od biologicznych – determinować ludzkie zachowania jedynie pośrednio. Pomostem między biologią a kulturą jest ludzki umysł i to on sprawił, że „znacząco większy procent repertuaru naszych zachowań – w porównaniu z innymi gatunkami – został przeniesiony na warstwę kulturową" (Virtual

[146] „On the one hand human beings are very much a product of evolution and each evolutionist products is special in its own right but we also have a particularly special version of "special" which involves the offloading of a much larger percentage of our behavioral repertoire to the cultural layer, and that interplay between our genomes, which are in many ways quite standard, and our cultural layer which is not matched by any other creature on Earth. That interplay is, one that has to be dealt with very carefully as you're thinking evolutionarily, but it is really the key to our being able to do what no other creature on Earth has ever been able to do." (*Harnessing Evolution - with Bret Weinstein | Virtual Futures Salon*, b.d.).

Futures, 2018). Dopiero wyposażone w umysły pojedyncze osobniki potrafią rozpoznać i wykorzystać wytwory kulturowe w podejmowanych przez siebie „zachowaniach wyższego rzędu", czyli złożonych działaniach intencjonalnych. Dlatego też stanowisko, że to właśnie wzajemne oddziaływanie między poziomem biologicznym a kulturowym jest „kluczowe dla naszego istnienia, gdyż umożliwia nam coś, czego żadne inne stworzenie na Ziemi nigdy nie było w stanie zrobić" (Virtual Futures, 2018), jest – w najlepszym razie – zbyt uproszczone. Nie uwzględnia ono tego, że zasadniczy wpływ na ludzkie zachowania ma pośredniczący między biologią a kulturą poziom procesów umysłowych. Powiedzieć można, że bez uwzględnienia „poziomu umysłowego" nie zrozumiemy ewolucji, jakiej uległy ludzkie zachowania. Przecież to treści wytworzone przez umysły zostały „wyładowane" do otoczenia i przybrały formę wytworów kulturowych (D. C. Dennett, 1997). Dlatego wyjaśnienie ludzkich zachowań wymaga w pierwszej kolejności określenia związków między nimi a procesami umysłowymi, a dopiero potem kulturowymi. Przekonanie o podstawowym znaczeniu procesów umysłowych dla złożonych ludzkich zachowań, a w szczególności dla działań intencjonalnych, leży u podstaw niniejszej pracy. Specyfika proponowanego tu podejścia polega na konstrukcji sekwencji modeli działania intencjonalnego. Każdy kolejny model jest coraz wierniejszym odwzorowaniem realnych działań intencjonalnych. W trakcie pierwszych prób stworzenia modelu złożonego działania intencjonalnego okazało się, że nie da się zbudować go, ograniczając się do wiedzy wypracowanej tylko w jednej dyscyplinie, stąd pojawiła się decyzja, żeby w jego konstrukcji posłużyć się językiem podejścia obliczeniowego. Język ten został jednak wykorzystany w taki sposób, aby za jego pomocą wyrazić wybrane idee filozoficzne, ustalenia psychologii intencji oraz neuronauki obliczeniowej. Przyjęto, iż dopiero uzgodnienie tych trzech „sposobów patrzenia" na działania intencjonalne pozwoli odsłonić ich złożoność oraz zasadnicze rodzaje związków między decydującymi o tych działaniach podsystemami.

Przystępując do konstrukcji modelu działania intencjonalnego w pracy wykorzystano idee wypracowane w następujących koncepcjach badawczych : teorii intencjonalności Searle'a (Searle, 1983), badaniach z obszaru psychologii intencji (Haggard, 2005; Libet, 2004; Patrick Haggard i in., 2002; Wegner, 2002) oraz obliczeniowych podstawach procesów decyzyjnych (M. Cichosz, 2010; P. Cichosz, 2007; Montague, 2006). Każda z tych teorii w odmienny sposób charakteryzuje czynniki, które są kluczowe dla zachowań ludzkich. Dane eksperymentalne, modele oraz koncepcje teoretyczne, które zostały

przedstawione i omówione w dysertacji, pokazują, że działania intencjonalne są wielowymiarowymi zjawiskami o złożonej dynamice. Integracja wiedzy zaczerpniętej z wymienionych koncepcji była ważnym celem pracy. Wymagało to zidentyfikowania kluczowych mechanizmów funkcjonowania działań intencjonalnych oraz wykorzystywanych do tego celu reprezentacji. Na tej podstawie zaproponowany został hierarchicznie zorganizowany model, który posłużył do wyjaśnienia najważniejszych cech działań intencjonalnych. Niewątpliwie najtrudniejszym do rozwiązania problemem poruszonym w pracy była kwestia wyjaśnienia charakterystycznego dla naszego gatunku cyklu rozwojowego dotyczącego działań intencjonalnych: od jego inicjalnego stanu, w pełni uwarunkowanego genetycznie, po złożone działania zdeterminowane kulturowo.

Struktura modelu przedstawionego w rozprawie jest rezultatem „współpracy" trzech podsystemów: (1) podsystemu hierarchicznego uczenia się ze wzmacnianiem z optymalizacją domenową (P-H-RL-OD), (2) podsystemu planowania i realizacji planów (P-PRP) oraz (3) podsystemu zarządzania siecią stanów intencjonalnych (P-ZSSI). Tak określony system – w zależności od poziomu zaawansowania wymienionych podsystemów – generuje różne formy zachowań. Początkowo, kiedy dominuje najstarszy ewolucyjnie podsystem uczenia się ze wzmacnianiem, zachowania są stosunkowo proste i ściśle skorelowane z biologicznymi, wrodzonymi potrzebami agenta. Z czasem wrodzone odruchy – na skutek uczenia się – przekształcane są w celowościowe zachowania wyższego poziomu. W tym okresie – poprzez eksplorację – dochodzi również do rozpoznania reguł określających funkcjonowanie środowiska (tzw. dyspozycje tła – wiedza-jak). Co szczególnie ważne, oprócz wyjaśnień neuronalno-behawioralno-introspekcyjnych, dysponujemy również obliczeniowym modelem tego podsystemu. Badania z obszaru uczenia maszynowego pozwoliły zidentyfikować i rozpoznać najważniejsze cechy algorytmów implementujących tę formę uczenia się. Zarówno teoretyczne analizy, jak i praktyczne zastosowania pokazują, że metoda uczenia się ze wzmacnianiem ma olbrzymi potencjał generalizujący, może ona skutecznie działać w złożonym i zmiennym środowisku, dlatego jest tak powszechna w świecie przyrody. Warto dodać, że obecnie algorytmy uczenia się ze wzmacnianiem stanowią fundament projektów dotyczących tzw. ogólnej (*general*) sztucznej inteligencji (Hassabis, 2017), która – w odróżnieniu od wąskiej (*narrow*) – może być przydatna w rozwiązywaniu całej klasy problemów, a nie tylko pojedynczych zadań. Optymalizacja funkcji wartości (V), za pomocą której wyznacza się strategię zachowań w algorytmach RL, nie jest jedynym zadaniem podsystemu uczenia się

ze wzmacnianiem. Innym ważnym skutkiem działania tego podsystemu jest tworzenie strumienia elementarnych warunków spełniania, czyli ustrukturyzowanych w związki przyczynowo-skutkowe reprezentacji aktywowanych wtedy, gdy agent wchodzi w różne interakcje ze środowiskiem. Na ich podstawie tworzona jest sieć najbardziej złożonych reprezentacji (stanów intencjonalnych), która umożliwia konstrukcję nowych typów zachowań, w szczególności zachowań kulturowych. Reprezentacjotwórcza funkcja metody uczenia się ze wzmacnianiem oparta jest na oryginalnej interpretacji teorii intencjonalności Johna Searle'a oraz na wybranych wynikach badań przeprowadzonych przez psychologów intencji, w szczególności na konstruktywistycznej koncepcji sprawstwa Daniela Wegnera oraz na korelacyjnej interpretacji intencji w działaniu zaproponowanej przez Patricka Haggarda.

Koniecznym dopełnieniem wymienionych podsystemów jest odpowiednio zaawansowany mechanizm planowania. W zaproponowanym modelu jest to niezależny podsystem, który z jednej strony wykorzystuje zasoby sieci stanów intencjonalnych, a z drugiej – dostarcza reprezentacji umożliwiających skrócenie procesu eksploracji w podsystemie uczenia się ze wzmacnianiem. Przyjęte rozwiązanie (wykorzystujące ideę prowizorycznych planów) stanowi kompromis polegający na połączeniu wyników krytycznej analizy Montague (wskazującej na ograniczone korzyści uzyskiwane ze szczegółowego planowania w środowisku naturalnym) z wynikami badań prowadzonych przez informatyków, którzy poszukują metod obliczeniowych pozwalających skrócić fazę eksploracji w algorytmach RL. Przyjęta koncepcja opiera się na hierarchicznym „podziale pracy". Podsystem planowania „specyfikuje" wysokopoziomowe, niezbędne do realizacji celu, reprezentacje (m.in. zachowania wysokiego poziomu – Z, nagrody – R, główne etapy realizacji celu (sekwencje zachowań Z, nagrody kształtujące) i przekazuje je na wejście podsystemowi uczenia się ze wzmacnianiem, który – na ich podstawie oraz na podstawie zdobytych wcześniej doświadczeń (zakodowanych w formie funkcji wartości V) – selekcjonuje zachowania. „Współpraca" wymienionych podsystemów umożliwia optymalizację osiąganych celów oraz wytwarzanie nowych form zachowań. Decyduje o tym szczególna właściwość naszego układu nerwowego, czyli zdolność do nadawania statusu nagrody wybranym, często abstrakcyjnym, reprezentacjom (por. hipoteza „nad-mocy" Reada Montague), które wyrażane są np. w chęci posiadania wyróżnionego statusu w grupie, w potrzebie realizacji dalekosiężnych planów czy w różnych przeżyciach religijnych. Uzyskanie tego typu zdolności domyka proces rozwoju całego układu.

Zdolność do prowizorycznego planowania, która opiera się na wiedzy zawartej w sieci stanów intencjonalnych dostępnej w formie językowej, jest najbardziej złożoną formą kontroli zachowań zarejestrowaną do tej pory w świecie przyrody.

Swoistym „efektem ubocznym" przyjętego w pracy podejścia badawczego są nieoczywiste reinterpretacje wybranych tez, pochodzących z wykorzystanych w pracy koncepcji. Do najważniejszych tego typu reinterpretacji zaliczyć można: (1) osłabienie hipotezy Montague dotyczącej małej użyteczności szczegółowego planowania, (2) przypisanie Wegnerowskiemu procesowi sprawstwa dodatkowej funkcji polegającej na rozszerzaniu i modyfikacji sieci stanów intencjonalnych, (3) odniesienie Searle'owskich warunków spełniania do reprezentacji generowanych przez mechanizm uczenia się ze wzmacnianiem, (4) zinterpretowanie dyspozycji tła jako złożonych zachowań wyższego rzędu. Głównym źródłem zaproponowanych zmian była potrzeba wyjaśnienia nie tylko złożonych działań intencjonalnych, ale również ich rozwojowego profilu, w szczególności – różnic między zachowaniami dzieci i osób dorosłych.

Należy podkreślić, że opracowanie bardziej adekwatnego modelu działań intencjonalnych stanie się możliwe dopiero wtedy, kiedy pojawią się jeszcze bardziej zaawansowane modele obliczeniowe mechanizmów organizacji zachowań. Wszystkie próby poprzedzające ten moment należy traktować jako mniej lub bardziej pożyteczne aproksymacje. Autor niniejszej pracy ma nadzieję, że zaproponowany w pracy model działań intencjonalnych pozwala dostrzec złożoność ich struktury, a w konsekwencji także i to, że ich wyjaśnienie wymaga integrowania wiedzy z różnych dyscyplin.

7 Bibliografia

Apperly, I. A., & Butterfill, S. A. (2009). Do Humans Have Two Systems to Track Beliefs and Belief-Like States? *Psychological Review*, *116*(4), 953–970. https://doi.org/10.1037/a0016923

Armstrong, J. (2015, czerwiec 17). *Los Angeles Review of Books*. Los Angeles Review of Books. https://lareviewofbooks.org/article/vision-science/

Asokan, A. (2016, wrzesień 24). *Brain against the machine*. The Hindu. http://www.thehindu.com/thread/technology/article9142676.ece

Barrett, L. F. (2016). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*.

Barrett, L. F. (2018). *Jak powstają emocje: Sekretne życie mózgu* (A. Jarosz, Tłum.). CeDeWu.

Bayne, T. (2006). *Phenomenology and the Feeling of Doing: Wegner on the Conscious Will*. The MIT Press.

Bekkering, H., Wohlschlager, A., & Gattis, M. (2000). Imitation of Gestures in Children is Goal-directed. *The Quarterly Journal of Experimental Psychology Section A*, *53*(1), 153–164. https://doi.org/10.1080/713755872

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, *74*(3), 183–200. https://doi.org/10.1037/h0024835

Bengio, Y. (2017). The Consciousness Prior. *arXiv:1709.08568 [cs, stat]*. http://arxiv.org/abs/1709.08568

Berridge, K. c, & Kringelbach, M. l. (2015). Pleasure Systems in the Brain. *Neuron*, *86*(3), 646–664. https://doi.org/10.1016/j.neuron.2015.02.018

Bewig, P. L. (2007, 21 października). *Streams*. https://srfi.schemers.org/srfi-41/srfi-41.html

Birkmayer, W., & Hornykiewicz, O. (1962). Der L-Dioxyphenylalanin (=L-DOPA)-Effekt beim Parkinson-Syndrom des Menschen: Zur Pathogenese und Behandlung der Parkinson-Akinese. *Archiv für Psychiatrie und Nervenkrankheiten*, *203*(5), 560–574. https://doi.org/10.1007/BF00343235

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology*, *76*, 198–211. https://doi.org/10.1016/j.jmp.2015.11.003

Borbone, G. (2011). Leszek Nowak and the idealizational approach to science. (Report). *Linguistic and Philosophical Investigations*, *10*, 125.

Brandon, R. N. (1978). Adaptation and evolutionary theory. *Studies in History and Philosophy of Science Part A*, *9*(3), 181–206. https://doi.org/10.1016/0039-3681(78)90005-5

Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. *Trends in Cognitive Sciences*, *16*(2), 106–113. https://doi.org/10.1016/j.tics.2011.12.010

Braver, T., S., & Cohen, J. D. (2000). On the Control of Control: The Role of Dopamine in Regulating Prefrontal Function and Working Memory. W S. Monsell & J. Driver (Red.), *Control of cognitive processes: Attention and performance XVIII*. MIT Press.

Brentano, F. C. (1999). *Psychologia z empirycznego punktu widzenia* (W. Galewicz, Tłum.). Wydawnictwo Naukowe PWN.

Bugg, J. M., Reisberg, D., Gallo, D. A., McDaniel, M. A., Wheeler, M. E., & Einstein, G. O. (2013). *Event-Based Prospective Remembering: An Integration of Prospective Memory and Cognitive Control Theories* (T. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195376746.013.0018

Cabanac, M. (1992). Pleasure: The common currency. *Journal of Theoretical Biology*, *155*(2), 173–200. https://doi.org/10.1016/S0022-5193(05)80594-6

Calvin, W. H. (1983). *The throwing madonna: Essays on the brain*. McGraw-Hill.

Carlson, S. M., Davis, A. C., & Leach, J. G. (2005). Less is more: Executive function and symbolic representation in preschool children. *Psychological Science*, *16*(8), 609–616. https://doi.org/10.1111/j.1467-9280.2005.01583.x

Carpenter, W. B. (1883). *Principles of mental physiology: With their applications to the training and discipline of the mind, and the study of its morbid conditions*. DAppleton. http://nrs.harvard.edu/urn-3:HUL.FIG:007441266

Caste and ecology in the social insects. (1979). *Acta Biotheoretica*, *28*(3), 234–235. https://doi.org/10.1007/BF00046355

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Chrisholm, R. M. (1966). Freedom and action. W K. Lehrer (Red.), *Freedom and determinism*. Random House.

Chrudzimski, A. (1995). Teoria Intencjonalności i umysłu Johna R. Searle'a. *Przeglad Filozoficzny - Nowa Seria*, *14*(2), 73–83.

Churchland, P. M. (1998). *On the contrary: Critical essays, 1987-1997*. MIT Press.

Churchland, P. S., & Sejnowski, T. J. (1992). *The computational brain*. MIT Press.

Cichosz, M. (2008). Znaczenie intencji dla moralnej oceny czynu – stanowisko Daniela Wegnera. *Analiza i Egzystencja*, 49–63.

Cichosz, M. (2010). Iluzja sprawczej funkcji intencji działania a mechanizm ustanawiania i osiągania celu. *Studia z Kognitywistyki i Filozofii Umysłu*, *4*(1). http://philpapers.org/rec/CICISF

Cichosz, P. (2007). *Systemy uczące się* (Wyd. 2). Wydawnictwa Naukowo-Techniczne.

Clark, A., & Toribio, J. (1994). Doing Without Representing? *Synthese*, 401–431.

Clarke, D. D., & Sokoloff, L. (1999). *Regulation of cerebral metabolic rate*. http://www.ncbi.nlm.nih.gov/books/NBK28194

Crick, F. (1997). *Zdumiewająca hipoteza czyli Nauka w poszukiwaniu duszy* (B. Chacińska-Abrahamowicz & M. Abrahamowicz, Tłum.). Prószyński i S-ka.

Crick, F., & Koch, C. (2008). Rama pojęciowa dla świadomości. *Formy aktywności umysłu. Ujęcia kognitywistyczne, t.1, Emocje, percepcja, świadomość. Wydawnictwo Naukowe PWN*.

Crutcher, R. (1994). Telling what we know: The use of verbal report methodologies in psychological research -- Protocol Analysis: Verbal Reports as Data (rev. ed.) by K. A. Ericsson and H. A. Simon. *Psychological Science, 5*(5), 241.

Damasio, A. R. (2011). *Błąd Kartezjusza: Emocje, rozum i ludzki mózg* (M. Karpiński, Tłum.; Wyd. 2 popr). Dom Wydawniczy „Rebis".

Davidson, D. (1973). *Essays on freedom of action* (T. Honderich, Red.). London, Boston, Routledge and Kegan Paul.

Davidson, D. (2001). *Freedom to Act*. Oxford University Press. https://doi.org/10.1093/0199246270.003.0004

Davidson, T. J., Kloosterman, F., & Wilson, M. A. (2009). Hippocampal Replay of Extended Experience. *Neuron*, *63*(4), 497–507. https://doi.org/10.1016/j.neuron.2009.07.027

Dayan, E., & Cohen, L. G. (2011). Neuroplasticity Subserving Motor Skill Learning. *Neuron, 72*(3), 443–454. https://doi.org/10.1016/j.neuron.2011.10.008

Hassabis, D. (2017). *DeepMind—Learning From First Principles—Artificial Intelligence* [NIPS]. https://www.youtube.com/watch?time_continue=1&v=DXNqYSNvnjA

Dennett, D. C. (1997). *Natura umysłów* (W. Turopolski, Tłum.). Wydawnictwo CIS.

Dennett, D., & Kinsbourne, M. (1992). Time and the observer: The where and when of consciousness in the brain. *The Behavioral and Brain Sciences, 15*(2), 183.

Descartes, R. (1958). *Medytacje o pierwszej filozofii* (K. Ajdukiewicz & M. Ajdukiewicz, Tłum.). Państwowe Wydawnictwo Naukowe.

Douglas, J., & Sutton, A. (1978). The development of speech and mental processes in a pair of twins: A case study. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *19*(1), 49–56.

Dretske, F. I. (2004). *Naturalizowanie umysłu* (Polska Akademia Nauk, Red.; B. Świątczak, Tłum.). Wydaw. Instytutu Filozofii i Socjologii PAN.

Drewery, A. (2000). Review of On the Contrary: Critical Essays, 1987-1997 [Review of *Review of On the Contrary: Critical Essays, 1987-1997*, P. M. Churchland & P. S. Churchland]. *The British Journal for the Philosophy of Science*, *51*(3), 507–511.

Duch, W. (2000). *Świadomość i dynamiczne modele działania mózgu*. Uniwersytet Mikołaja Kopernika.

Falk, D. (2005). Great Eureka Moments in History. *University of Toronto Magazine*. https://magazine.utoronto.ca/research-ideas/culture-society/great-eureka-moments-in-history-famous-inspirational-moments/

Felski, R. (2011). Critique and the Hermeneutics of Suspicion. *M/C Journal*, *15*(1). http://journal.media-culture.org.au/index.php/mcjournal/article/view/431

Festinger, L. (1957). *A theory of cognitive dissonance.* Row, Peterson.

Flanagan, J. R., & Johansson, R. S. (2003). Action plans used in action observation. *Nature*, *424*(6950), 769.

Frege, G. (1977). *Pisma semantyczne* (B. Wolniewicz, Tłum.). Państwowe Wydawnictwo Naukowe.

Fried, I., Katz, A., Mccarthy, G., Sass, K. J., Williamson, P., Spencer, S. S., & Spencer, D. D. (1991). Functional organization of human supplementary motor cortex studied by electrical stimulation. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *11*(11), 3656–3666.

Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, *16*(9), 1325–1352. https://doi.org/10.1016/j.neunet.2003.06.005

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K. (2017, maj 18). *Consciousness is not a thing, but a process of inference*. Aeon Essays. https://aeon.co/essays/consciousness-is-not-a-thing-but-a-process-of-inference

Frith, C. (2012). Explaining delusions of control: The comparator model 20 years on. *Consciousness and Cognition*, *21*(1), 52–54. https://doi.org/10.1016/j.concog.2011.06.010

Fromm, E. (2006). *Rewizja psychoanalizy* (K. Pospiszyl & R. Saciuk, Red.; Wyd. 2 zm). Wydawnictwo Naukowe PWN.

Gallistel, C. R., & King, A. P. (2011). *Memory and the Computational Brain: Why Cognitive Science will Transform Neuroscience*. John Wiley & Sons.

Gazzaniga, M. S. (1978). *The integrated mind*. Plenum Press.

Gershman, S. J. (2019). What does the free energy principle tell us about the brain? *arXiv:1901.07945 [q-bio]*. http://arxiv.org/abs/1901.07945

Goel, V., Gold, B., Kapur, S., & Houle, S. (1998). Neuroanatomical correlates of human reasoning. *Journal of Cognitive Neuroscience*, *10*(3), 293–302.

Goertzel, B., & Pennachin, C. (2007). *Artificial General Intelligence*. Springer Berlin Heidelberg.

Gomes, G. (1998). The Timing of Conscious Experience: A Critical Review and Reinterpretation of Libet's Research. *Consciousness and Cognition*, *7*(4), 559–595. https://doi.org/10.1006/ccog.1998.0332

Gould, S. J., & Lewontin, R. C. (1979). The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. *Proceedings of the Royal Society of London. Series B, Biological Sciences (1934-1990)*, *205*(1161), 581–598. https://doi.org/10.1098/rspb.1979.0086

Graham, G. (2017). Behaviorism. W E. N. Zalta (Red.), *The Stanford Encyclopedia of Philosophy* (Spring 2017). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2017/entries/behaviorism/

Green, M. (2017). Speech Acts. W E. N. Zalta (Red.), *The Stanford Encyclopedia of Philosophy* (Winter 2017). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2017/entries/speech-acts/

Groman, S. M., Massi, B., Mathias, S. R., Lee, D., & Taylor, J. R. (2019). Model-Free and Model-Based Influences in Addiction-Related Behaviors. *Biological Psychiatry*, *85*(11), 936–945. https://doi.org/10.1016/j.biopsych.2018.12.017

Grześ, M. (2010). *Improving Exploration in Reinforcement Learning through Domain Knowledge and Parameter Analysis* [Thesis, University of York]. http://etheses.whiterose.ac.uk/936/

Gu, X., Lohrenz, T., Salas, R., Baldwin, P. R., Soltani, A., Kirk, U., Cinciripini, P. M., & Montague, R. P. (2015). Belief about nicotine selectively modulates value and reward prediction error signals in smokers. *Proceedings of the National Academy of Sciences*, *112*(8), 2539. https://doi.org/10.1073/pnas.1416639112

Gut, A. (2015). Czy reprezentacje zwierząt są nieprzezroczyste? *Przegląd Filozoficzny – Nowa Seria*, 371–382.

Haggard, P., Aschersleben, G., Gehrke, J., & Prinz, W. (2002). Action, binding and awareness. *Undefined*. https://www.semanticscholar.org/paper/Action%2C-binding-and-awareness-Haggard-Aschersleben/6cc5b92ba02d9ec97627550a46854e8afebc0760

Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience*, *5*(4), 382. https://doi.org/10.1038/nn827

Haggard, P. (2005). Conscious intention and motor cognition. *Trends in Cognitive Sciences*, *9*(6), 290–295. https://doi.org/10.1016/j.tics.2005.04.012

Haggard, P. (2008). Human volition: Towards a neuroscience of will.(Report). *Nature Reviews Neuroscience*, *9*(12), 934.

Haggard, P. (2012a). *Conscious intention and brain activity*. http://www.ingentaconnect.com/content/imp/jcs/2001/00000008/00000011/1238

Haggard, P. (2012b). *Watch „Who's in Control: Patrick Haggard at TEDxHelvetia" Video at TEDxTalks*. https://www.youtube.com/watch?v=knvGvWghRIE

Harari, Y. N. (2018). *Homo deus: Krótka historia jutra* (M. Romanek, Tłum.; Wydanie pierwsze). Wydawnictwo Literackie.

Harris, S. (2012). *Free will* (1st Free Press trade pbk. ed.). Free Press.

Harwas-Napierała, B., & Trempała, J. (Red.). (2004). *Psychologia rozwoju człowieka. T. 2, Charakterystyka okresów życia człowieka* (Wyd. 3). Wydaw. Naukowe PWN.

Heisters, D. (2011). Parkinson's: Symptoms, treatments and research. *British Journal of Nursing (Mark Allen Publishing)*, *20*(9), 548–554. https://doi.org/10.12968/bjon.2011.20.9.548

Herling-Grudziński, G. (1990). *Dziennik pisany nocą 1973-1979*. Res Publica.

Hoffman, R. E. (1986). Verbal hallucinations and language production processes in schizophrenia. *Behavioral and Brain Sciences*, *9*(3), 503–517. https://doi.org/10.1017/S0140525X00046781

Hobbes, T. (2009). *Lewiatan czyli Materia, forma i władza państwa kościelnego i świeckiego* (C. Znamierowski, Tłum.). Fundacja Aletheia.

Honderich, T. (2001). *Ile mamy wolności?: Problem determinizmu* (A. Florek, Tłum.). Zysk i S-ka.

Hurlburt, R. T., & Heavey, C. L. (2001). Telling what we know: Describing inner experience. *Trends in Cognitive Sciences*, *5*(9), 400–403. https://doi.org/10.1016/S1364-6613(00)01724-1

Hyperparameter. (2019, maja 17). DeepAI. https://deepai.org/machine-learning-glossary-and-terms/hyperparameter

Inoue, S., & Matsuzawa, T. (2007a). Working memory of numerals in chimpanzees. *Current Biology*, *17*(23), R1004–R1005. https://doi.org/10.1016/j.cub.2007.10.027

Jackson, F. (1982). Epiphenomenal Qualia. *The Philosophical Quarterly*, *32*(127), 127–136. https://doi.org/10.2307/2960077

Jacob, P. (2014). Intentionality. W E. N. Zalta (Red.), *The Stanford Encyclopedia of Philosophy* (Winter 2014). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2014/entries/intentionality/

James, W. (1950). *The principles of Psychology. Vol. 1*. Dover Publications.

Jodzio, K. (2008). *Neuropsychologia intencjonalnego działania: Koncepcje funkcji wykonawczych*. Wydawnictwo Naukowe Scholar.

Johnson-Laird, P. N. (2006). *How we reason*. Oxford University Press. http://nrs.harvard.edu/urn-3:hul.ebookbatch.GEN_batch:EDZ000002399020160630

Jurgielewicz-Delegacz, E. (2019). Ewolucja odpowiedzialności nieletnich na przestrzeni lat. *Studia Prawnoustrojowe*, *44*, 171–186. https://doi.org/10.31648/sp.4902

Kahneman, D., & Tversky, A. (2012). *Pułapki myślenia: O myśleniu szybkim i wolnym* (P. Szymczak, Tłum.). Media Rodzina.

Kaplan, R., & Friston, K. J. (2018). Planning and navigation as active inference. *Biological Cybernetics*, *112*(4), 323–343. https://doi.org/10.1007/s00422-018-0753-2

Klatzky, R. L. (1998). Allocentric and Egocentric Spatial Representations: Definitions, Distinctions, and Interconnections. W C. Freksa, C. Habel, & K. F. Wender (Red.), *Spatial Cognition: An Interdisciplinary Approach to Representing and Processing Spatial Knowledge* (s. 1–17). Springer. https://doi.org/10.1007/3-540-69342-4_1

Kim, J. (1995). Mental Causation in Searle's „Biological Naturalism". *Philosophy and Phenomenological Research*, *55*(1), 189–194. https://doi.org/10.2307/2108318

Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *21*(16), RC159.

Korbak, T. (2019). Computational enactivism under the free energy principle. *Synthese*. https://doi.org/10.1007/s11229-019-02243-4

Kornhuber, H., & Deecke, L. (1965). Hirnpotentialänderungen bei Willkürbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale. *Pflüger's Archiv für die gesamte Physiologie des Menschen und der Tiere*, *284*(1), 1–17. https://doi.org/10.1007/BF00412364

Korpikiewicz, H. (2017). Instynkt – naśladownictwo – myślenie. Jak się uczą zwierzęta. *Filozofia Publiczna i Edukacja Demokratyczna*, *6*(1), 129–150. https://doi.org/10.14746/fped.2017.6.1.8

Kripke, S. (2001). *Nazywanie a konieczność* (B. Chwedeńczuk, Tłum.; Dwuwydziałowa Biblioteka Nauk Społecznych ul. Szamarzewskiego 91 NA-74553). Fundacja Aletheia. http://han.amu.edu.pl/han/academicsearchcomplete/search.ebscohost.com/login.aspx?direct=true&db=cat07374a&AN=uamp.310966&lang=pl&site=eds-live&scope=site

Kruel, A. (2013, lipiec 13). *Narrow vs. General Artificial Intelligence*. http://kruel.co/2013/07/13/narrow-vs-general-artificial-intelligence/

Kuhn, M. (2013). *Applied Predictive Modeling* (1st ed. 2013.). Springer New York : Imprint: Springer.

Kulkarni, T. D., Narasimhan, K. R., Saeedi, A., & Tenenbaum, J. B. (2016). *Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation*. http://arxiv.org/abs/1604.06057

Kulke, L., Johannsen, J., & Rakoczy, H. (2019). Why can some implicit Theory of Mind tasks be replicated and others cannot? A test of mentalizing versus submentalizing accounts. *PloS One*, *14*(3), e0213772–e0213772. https://doi.org/10.1371/journal.pone.0213772

Lasota, M., & Grenda, B. (2017). *Arena samobójców: Wybrane aspekty terroryzmu i terroryzmu samobójczego*. Wydawnictwo Akademii Sztuki Wojennej.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building Machines That Learn and Think Like People. *arXiv:1604.00289 [cs, stat]*. http://arxiv.org/abs/1604.00289

Lakshminarayanan, A. S., Krishnamurthy, R., Kumar, P., & Ravindran, B. (2016). *Option Discovery in Hierarchical Reinforcement Learning using Spatio-Temporal Clustering*. http://arxiv.org/abs/1605.05359

Lau, H. C., Rogers, R. D., Haggard, P., & Passingham, R. E. (2004). Attention to intention. (Reports; brain chemistry). *Science*, *303*(5661), 1208.

Lau, H. C., Rogers, R. D., & Passingham, R. E. (2007). Manipulating the Experienced Onset of Intention after Action Execution. *Journal of Cognitive Neuroscience*, *19*(1), 81–90. https://doi.org/10.1162/jocn.2007.19.1.81

Lebiere, C., & Lee, F. J. (2002). Intention superiority effect: A context-switching account. *Cognitive Systems Research*, *3*(1), 57–65. https://doi.org/10.1016/S1389-0417(01)00044-4

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current opinion in neurobiology*, *22*(6), 1027–1038. https://doi.org/10.1016/j.conb.2012.06.001

Li, H., Kulik, L., & Ramamohanarao, K. (2015). Robust inferences of travel paths from GPS trajectories. *International Journal of Geographical Information Science*, *29*(12), 2194–2222. https://doi.org/10.1080/13658816.2015.1072202

Libet, B. (2004). *Mind time: The temporal factor in consciousness*. Harvard University Press.

Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain (London, England : 1878)*, *106 (Pt 3)*(3), 623–642. https://doi.org/10.1093/brain/106.3.623

Libet, B., Wright, E. W., & Gleason, C. A. (1983). Preparation- or intention-to-act, in relation to pre-event potentials recorded at the vertex. *Electroencephalography and*

Clinical Neurophysiology, *56*(4), 367–372. https://doi.org/10.1016/0013-4694(83)90262-6

Loughlin, V. (2017). Jakob Hohwy: The predictive mind. *Phenomenology and the Cognitive Sciences*, *16*(4), 753–758. https://doi.org/10.1007/s11097-016-9479-6

Luriĭa, A. R. (1959). *Speech and the development of mental processes in the child; an experimental investigation*. Staples Press.

Malinowska, J. K. (2016). Cultural neuroscience and the category of race: The case of the other-race effect. *Synthese*, *193*(12), 3865–3887. https://doi.org/10.1007/s11229-016-1108-y

Marr, D. (2010). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. The MIT Press. https://doi.org/10.7551/mitpress/9780262514620.001.0001

Marsden, C. A. (2006). Dopamine: The rewarding years. *British Journal of Pharmacology*, *147 Suppl 1*, S136-44.

McGrew, W. (2003). Culture in nonhuman primates? *Annual Review of Anthropology*, *27*, 301–328. https://doi.org/10.1146/annurev.anthro.27.1.301

McClure, S. M., Li, J., Tomlin, D., Cypert, K. S., Montague, L. M., & Montague, P. R. (2004). Neural Correlates of Behavioral Preference for Culturally Familiar Drinks. *Neuron*, *44*(2), 379–387. https://doi.org/10.1016/j.neuron.2004.09.019

Mead, M. (1932). An Investigation of the Thought of Primitive Children, with Special Reference to Animism. *The Journal of the Royal Anthropological Institute of Great Britain and Ireland*, *62*, 173–190. https://doi.org/10.2307/2843884

Meijers, A. W. M. (2000). Mental Causation and Searle's Impossible Conception of Unconscious Intentionality. *International Journal of Philosophical Studies : IJPS*, *8*(2), 155–170. https://doi.org/10.1080/09672550050083974

Mele, A. R. (2009). *Effective intentions: The power of conscious will*. Oxford University Press. http://nrs.harvard.edu/urn-3:hul.ebookbatch.OXSCH_batch:osouk9780195384260

Miall, R. C., & Wolpert, D. M. (1996). Forward Models for Physiological Motor Control. *Neural Networks*, *9*(8), 1265–1279. https://doi.org/10.1016/S0893-6080(96)00035-4

Millidge, B., Seth, A., & Buckley, C. (2021). *Predictive Coding: A Theoretical and Experimental Review*.

Millikan, R. G. (2009). *Biosemantics*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199262618.003.0024

Miłkowski, M. (2009). Jak wyróżniać moduły umysłowe? Problemy ze specjalizacją i konfirmacją. *Studia z Kognitywistyki i Filozofii Umysłu*, *3*. https://philpapers.org/rec/MIKJWM

Miłkowski, M. (2015, 16 grudnia). *Rola filozofii w kognitywistyce i kognitywistyki w filozofii*. http://marcinmilkowski.pl/downloads/fu_II/wyk15.pptx

Model. (2016). W *Wikipedia, wolna encyklopedia*. https://pl.wikipedia.org/w/index.php?title=Model&oldid=46014786

Montague, P. R., Berns, G. S., Cohen, J. D., Mcclure, S. M., Pagnoni, G., Dhamala, M., Wiest, M. C., Karpov, I., King, R. D., Apple, N., & Fisher, R. E. (2002). Hyperscanning: Simultaneous fMRI during Linked Social Interactions. *Neuroimage*, *16*(4), 1159–1164. https://doi.org/10.1006/nimg.2002.1150

Montague, R. (2006). *Why choose this book?: How we make decisions*. Dutton.

Mülling, K., Kober, J., Kroemer, O., & Peters, J. (2013). Learning to select and generalize striking movements in robot table tennis. *The International Journal of Robotics Research*, *32*(3), 263–279. https://doi.org/10.1177/0278364912472380

Nagel, T. (2012). *What is it like to be a bat?* Cambridge University Press.

Nęcka, E., Orzechowski, J., & Szymura, B. (2006). *Psychologia poznawcza*. Academica Wydawnictwo SWSP : Wydawnictwo Naukowe PWN.

Niedźwieńska, A. (2013). *Pamięć prospektywna: Geneza, mechanizmy, deficyty*. Wydawnictwo Akademickie „Sedno".

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

Nowak, L. (1977). *Wstęp do idealizacyjnej teorii nauki*. Państwowe Wydawnictwo Naukowe.

O'Connor, T. (2020). Emergent Properties. W E. N. Zalta (Red.), *The Stanford Encyclopedia of Philosophy* (Fall 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/properties-emergent/

Olds, J. (1958). Self-stimulation of the brain; its use to study local effects of hunger, sex, and drugs. *Science (New York, N.Y.)*, *127*(3294), 315–324.

O'Neill, J., Pleydell-Bouverie, B., Dupret, D., & Csicsvari, J. (2010). Play it again: Reactivation of waking experience and memory. *Trends in Neurosciences*, *33*(5), 220–229. https://doi.org/10.1016/j.tins.2010.01.006

O'Reilly, R. C., Braver, T. S., & Cohen, J. D. (1999). A Biologically-Based Computational Model of Working Memory. W A. Miyake & P. Shah (Red.), *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge University Press.

Oster, G. F., & Wilson, E. O. (1978). *Caste and ecology in the social insects* (T. 12). Princeton University Press.

Oster, G. F. (1978). *Caste and ecology in the social insects*. Princeton University Press.

Otto, R. (2000). *Mistyka Wschodu i Zachodu: Analogie i różnice wyjaśniające jej istotę* (T. Duliński, Tłum.). KR.

Paulson, S. (2015, 23 kwietnia). *Ingenious: David Krakauer*. Nautilus. http://nautil.us/issue/23/dominoes/ingenious-david-krakauer

Parrila, R. K., Das, J. P., & Dash, U. N. (1996). Development of planning and its relation to other cognitive processes. *Journal of Applied Developmental Psychology*, *17*(4), 597–624. https://doi.org/10.1016/S0193-3973(96)90018-0

Penfield, W. (1975). *The Mystery of the Mind*. Princeton University Press.

Piaget, J. (1966). *Narodziny inteligencji dziecka* (M. Przetacznik-Gierowska, Tłum.). Państwowe Wydawnictwo Naukowe.

Pisella, L., Gréa, H., Tilikete, C., Vighetto, A., Desmurget, Rode, G., Boisson, D., & Rossetti, Y. (2000). An 'automatic pilot' for the hand in human posterior parietal cortex: Toward reinterpreting optic ataxia. *Nature Neuroscience*, *3*(7), 729. https://doi.org/10.1038/76694

PMBOK Guide and Standards | Project Management Institute. (2018, 7 stycznia). https://www.pmi.org/pmbok-guide-standards

Pockett, S., Banks, W. P., & Gallagher, S. (2006). *Does consciousness cause behavior?* MIT Press.

Pockett, S., & Miller, A. (2007). The rotating spot method of timing subjective events. *Consciousness and Cognition*, *16*(2), 241–254. https://doi.org/10.1016/j.concog.2006.09.002

Poczobut, R. (2009). *Między redukcją a emergencją: Spór o miejsce umysłu w świecie fizycznym*. Wydawnictwo Uniwersytetu Wrocławskiego.

Popper, K. (2014). Conjectures and Refutations: The Growth of Scientific Knowledge. Routledge.

Popper, K. R., & Hudson, G. E. (1963). Conjectures and Refutations. *Physics Today*, *16*(11), 80–82. https://doi.org/10.1063/1.3050617

Prinz, W. (1987). *Ideo-Motor Action* (Przez H. Heuer & A. Sanders; s. 61–90). Routledge. https://doi.org/10.4324/978131562799-11

Rangel, A., Camerer, C., & Montague, R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*(7), 545. https://doi.org/10.1038/nrn2357

Rakowska, J. M. (2005). *Skuteczność psychoterapii: Przegląd badań*. Wydawnictwo Naukowe „Scholar".

Reuter, M. (2014). Dziecięca teoria umysłu a rozwój funkcji wykonawczych. *Przegląd Filozoficzno-Literacki*, *0*(2 (39)), 189–203.

Richard Feynman cytaty: 18 cytatów i aforyzmów Richarda Feynmana. (2016, 25 października). cytatybaza.pl. http://cytatybaza.pl/autorzy/richard-feynman.html

Rosen, C. (2008). The Myth of Multitasking. *New Atlantis (Washington, D.C.)*, *20*, 105–110.

Roskies, A. (2006). Neuroscientific challenges to free will and responsibility. *Trends in Cognitive Sciences*, *10*(9), 419–423. https://doi.org/10.1016/j.tics.2006.07.011

Rousseau, J. J. (1930). *Emil czyli O wychowaniu* (W. Husarski, Tłum.). Naukowe Towarzystwo Pedagogiczne.

Sacks, O. (1999). *Antropolog na Marsie* (Piotr Amsterdamsk, Aleksander Radomski, Barbara Lindenberg, & Beata Maciejewska, Tłum.). Zysk i S-ka Wydawnictwo.

Sadowski, B. (2012). *Biologiczne mechanizmy zachowania się ludzi i zwierząt* (Wyd. 3, 4 dodr). Wydawnictwo Naukowe PWN.

Schlosser, M. (2019). Agency. W E. N. Zalta (Red.), *The Stanford Encyclopedia of Philosophy* (Winter 2019). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/agency/

Schroedl, S., Wagstaff, K., Rogers, S., Langley, P., & Wilson, C. (2004). Mining GPS Traces for Map Refinement. *Data Mining and Knowledge Discovery*, *9*(1), 59–87. https://doi.org/10.1023/B:DAMI.0000026904.74892.89

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *13*(3), 900–913.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward.(Special Section: Cognitive Neuroscience)(Cover Story). *Science*, *275*(5306), 1593.

Schultz, W., & Romo, R. (1990). Dopamine neurons of the monkey midbrain: Contingencies of responses to stimuli eliciting immediate behavioral reactions. *Journal of Neurophysiology*, *63*(3), 607–624.

Schurger, A., Sitt, J. D., & Dehaene, S. (2012). An accumulator model for spontaneous neural activity prior to self-initiated movement. *Proceedings of the National Academy of Sciences*, *109*(42), E2904–E2913. https://doi.org/10.1073/pnas.1210467109

Searle, J. R. (1983). *Intentionality, an essay in the philosophy of mind*. Cambridge University Press.

Searle, J. R. (1992). *The rediscovery of the mind*. MIT Press.

Searle, J. R. (1995). Consciousness, the Brain and the Connection Principle: A Reply. *Philosophy and Phenomenological Research*, *55*(1), 217–232. https://doi.org/10.2307/2108322

Searle, J. R. (1995). *The construction of social reality*. Free Press.

Searle, J. R. (2000). Mental Causation, Conscious and Unconscious: A Reply to Anthonie Meijers. *International Journal of Philosophical Studies : IJPS*, *8*(2), 171–177. https://doi.org/10.1080/09672550050083983

Searle, J. R. (2001). *Rationality in action*. MIT Press.

Searle, J. R. (2007). *Biological Naturalism* (s. 325–334). Blackwell Publishing. https://doi.org/10.1002/9780470751466.ch26

Searle, J. R. (2008a). *Philosophy in a New Century: Selected Essays*. Cambridge University Press. https://doi.org/10.1017/CBO9780511812859

Searle, J. R. (Red.). (2008b). Twenty-one years in the Chinese Room. W *Philosophy in a New Century: Selected Essays* (s. 67–85). Cambridge University Press. https://doi.org/10.1017/CBO9780511812859.006

Searle, J. R. (Red.). (2008c). Why I am not a property dualist. W *Philosophy in a New Century: Selected Essays* (s. 152–160). Cambridge University Press. https://doi.org/10.1017/CBO9780511812859.010

Searle, J. R. (2010a). *Consciousness and the Problem of Free Will*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195389760.003.0008

Searle, J. R. (2010b). *Umysł: Krótkie wprowadzenie* (J. Karłowski, Tłum.). Dom Wydawniczy „Rebis".

Searle, J. R. (2011). *Lecture 1: Cartesian Dualism, Mind-Body Problem, Perception | CosmoLearning Philosophy*. CosmoLearning. https://cosmolearning.org/video-lectures/cartesian-dualism-mind-body-problem-perception/

Searle, J. R. (2013). *John Searle: Świadomość—Wspólny ludzki stan. | TED Talk Subtitles and Transcript | TED.com*. https://www.ted.com/talks/john_searle_our_shared_condition_consciousness/transcript?language=pl

Searle, J. R. (2015). *Seeing things as they are: A theory of perception*. Oxford University Press.

Segerdahl, P., Fields, W., & Savage-Rumbaugh, E. S. (2005). *Kanzi's Primal Language: The Cultural Initiation of Primates into Language*. Palgrave Macmillan Limited.

Shaffer, D. R. (2010). *Developmental psychology: Childhood and adolescence* (8th ed.). Wadsworth, Cengage Learning.

Shear, J. (1997). *Explaining consciousness: The „hard problem"*. MIT Press.

Shephard, E., Jackson, G. M., & Groom, M. J. (2014). Learning and altering behaviours by reinforcement: Neurocognitive differences between children and adults. *Developmental Cognitive Neuroscience*, *7*, 94–105. https://doi.org/10.1016/j.dcn.2013.12.001

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Silver, D. (2015, marzec 5). *Value Function Approximation*. https://www.davidsilver.uk/wp-content/uploads/2020/03/FA.pdf

Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence*, *299*, 103535. https://doi.org/10.1016/j.artint.2021.103535

Singh, S., Barto, A. G., & Chentanez, N. (2005). *Intrinsically Motivated Reinforcement Learning*. http://handle.dtic.mil/100.2/ADA440280

Sirigu, A., Daprati, E., Ciancia, S., Giraux, P., Nighoghossian, N., Posada, A., & Haggard, P. (2003). Altered awareness of voluntary action after damage to the parietal cortex. *Nature Neuroscience*, *7*(1), 80. https://doi.org/10.1038/nn1160

Smith, B. (2003). *John Searle*. Cambridge University Press.

Soon, C. S., Brass, M., Heinze, H.-J., & Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, *11*(5), 543. https://doi.org/10.1038/nn.2112

Stahl, S. M. (2009a). *Podstawy psychofarmakologii: Teoria i praktyka. T. 1* (K. Grabowski, Tłum.). Via Medica.

Stahl, S. M. (2009b). *Podstawy psychofarmakologii: Teoria i praktyka. T. 2* (K. Grabowski, Tłum.). Via Medica.

STRIPS. (2015). W *Wikipedia*. https://en.wikipedia.org/w/index.php?title=STRIPS&oldid=649724191

Stuss, D. T., & Knight, R. T. (2002). *Principles of frontal lobe function*. Oxford University Press. http://nrs.harvard.edu/urn-3:hul.ebookbatch.OXSCH_batch:osouk9780195134971

Sutton, R. S. (1998). *Reinforcement learning: An introduction*. MIT Press. http://nrs.harvard.edu/urn-3:hul.ebook:EBSCO_1094

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*(1), 181–211. https://doi.org/10.1016/S0004-3702(99)00052-1

Sztompka, P., & Konieczny, J. (2005). *Socjologia zmian społecznych*. Znak.

Tatarkiewicz, W. (1995). *Historia filozofii. T. 3: Filozofia XIX wieku i współczesna* (Wyd. 14). Wydaw. Naukowe PWN.

Thagard, P. (2020). Cognitive Science. W E. N. Zalta (Red.), *The Stanford Encyclopedia of Philosophy* (Winter 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/cognitive-science/

Trojan, M. (2013). *Na tropie zwierzęcego umysłu*. Wydawnictwo Naukowe Scholar.

Van Dijk, S. (2003). *Reinforcement Learning*. Artificial Intelligence Institute, University of Groningen, The Netherlands. http://www.ai.rug.nl/ki2/slides/ki2-s11-reinforcement-learning.ppt

Virtual Futures. (2018, 31 marca). *Harnessing Evolution—With Bret Weinstein | Virtual Futures Salon*. https://www.youtube.com/watch?v=nOMLdefHGA8

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. https://doi.org/10.1038/nature14236

Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001). Constrained K-means Clustering with Background Knowledge. *Proceedings of the Eighteenth International Conference on Machine Learning*, 577–584. http://dl.acm.org/citation.cfm?id=645530.655669

Wegner, D. M. (2002). *The illusion of conscious will*. MIT Press.

Wegner, D. M., & Wheatley, T. (1999). Apparent Mental Causation: Sources of the Experience of Will. *American Psychologist*, *54*(7), 480.

Węcławski, T. (1995). *Wspólny świat religii*. Wydaw. Znak.

Wightman, R. M. (2006). Detection technologies. Probing cellular chemistry in biological systems with microelectrodes. *Science (New York, N.Y.)*, *311*(5767), 1570–1574.

Wilson, E. O. (1988). *O naturze ludzkiej* (B. Szacka, Tłum.). Państ. Instytut Wydawniczy.

Wilson, E. O. (2002). *Konsiliencja: Jedność wiedzy* (J. Mikos, Tłum.). Wydaw. Zysk i S-ka.

Wise, R. A. (2002). Brain Reward Circuitry: Insights from Unsensed Incentives. *Neuron*, *36*(2), 229–240. https://doi.org/10.1016/S0896-6273(02)00965-0

Yao, H., Szepesvari, C., Sutton, R. S., Modayil, J., & Bhatnagar, S. (2014). Universal Option Models. W Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Red.), *Advances in Neural Information Processing Systems 27* (s. 990–998). Curran Associates, Inc. http://papers.nips.cc/paper/5590-universal-option-models.pdf

Yoo, J. (2007). Mental Causation. W *Internet Encyclopedia of Philosophy*.

Zhao, D., Haitao Wang, Kun Shao, & Zhu, Y. (2016). Deep reinforcement learning with experience replay based on SARSA. *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–6. https://doi.org/10.1109/SSCI.2016.7849837

8 Lista diagramów

Diagram 1. Zintegrowany model działań intencjonalnych (ZMDI) – ujęcie wysokopoziomowe. 38

Diagram 2. Relacja między siecią stanów intencjonalnych a dyspozycjami tła. 63

Diagram 3. Schemat przebiegu działania intencjonalnego wg Searle'a (Searle, 1983, s. 98). 65

Diagram 4. Intencja w działaniu z perspektywy problemu umysł-ciało w ujęciu Searle'a. 89

Diagram 5. Schemat algorytmu uczenia ze wzmacnianiem (za: Van Dijk 2008). 101

Diagram 6 Typy reprezentacji w modelu OBC. 126

Diagram 7 Model działania dowolnego wg Daniela Wegnera. 165

Diagram 8. Model kontroli ruchowej (por. Haggard, 2005). 176

Diagram 9. Główne składowe działania intencjonalnego wg. Searle'a. 190

Diagram 10. Model działania intencjonalnego uwzględniający wyłącznie powiązania niskopoziomowe wersja 1.0. 224

Diagram 11. Model działania intencjonalnego z podsystemem kontroli celów i podsystemem projektowania ich zmiany wersja 1.1. 228

Diagram 12. Model 1.2 działania intencjonalnego z kreatorem nowych typów zachowań oraz ewaluatorem stanów umysłowych jako nowego typu nagród. 235

Diagram 13. Model działania intencjonalnego z podsystemem zarządzania siecią stanów intencjonalnych (P-ZSSI). 254

Diagram 14. Model działania intencjonalnego z podsystemem zarządzania siecią stanów intencjonalnych (P-ZSSI). 275

9 Lista rysunków

Rys. 1 Rozkład nagród w przykładowym środowisku robota. 104
Rys. 2 Początkowy stan funkcji wartości V. 104
Rys. 3 Stan środowiska po przejściu ze stanu s1 do s2. 105
Rys. 4 Stan środowiska po przejściu całego korytarza - aż do stanu s7. 105
Rys. 5 Stan środowiska po cofnięciu się robota ze stanu s7 do s1. 106
Rys. 6 Wartości V dla stanów środowiska po wielokrotnym przejściu korytarza. 106

10 Lista ilustracji

Ilustracja 1. Układ dla eksperymentu I-Spy (Wegner, 2002, s. 75). 171

11 Legenda symboli

Symbole użyte w Modelu 1.0:

- z_t – zachowanie zrealizowane przez agenta w chwili t; za wybór zachowania z_t odpowiada podsystem P-RL;
- b_{t+1} – bodźce odebrane przez agenta w chwili t+1, wygenerowane przez środowisko, będące w stanie s_t w związku z zachowaniem z_t, np. b_{t+1} to sygnały świetlne docierające do siatkówki oka odbierane w związku z wykonaniem określonego ruchu głowy;
- o_{t+1} – obserwacja o_{t+1} jest reprezentacją utworzoną przez podsystem sensoryczno-ewaluacyjny (P-SE); odnosi się do stanu środowiska w chwili 't+1'; podstawą do utworzenia tego typu reprezentacji są bodźce b_{t+1}, które są reakcją środowiska na zachowanie z_t, którego realizacji podjął się agent, np. agent tworzy odpowiednią reprezentację percepcyjną na podstawie pobudzenia układu wzrokowego, a następnie rozpoznaje – na podstawie określnych cech tej reprezentacji – że znalazł się w stanie s_t;
- r_{t+1} – nagroda r_{t+1} to pochodząca ze środowiska natychmiastowa zwrotna informacja wartościująca, oceniająca stan 's_t' z perspektywy realizowanego celu; za utworzenie tego typu reprezentacji odpowiedzialny jest określony moduł w podsystemie sensorycznym (P-SE), który – na podstawie docierających do agenta pobudzeń sensorycznych – wyznacza ich bieżącą wartość, np. wartość pożywienia określona jest na bazie sygnałów pochodzących z układu węchowo-smakowego.

Symbole dodane do Modelu 1.1:

- $[b_ś\ldots]_{t+1}$, $[b_h\ldots]_{t+1}$ – bodźce pochodzące ze środowiska zewnętrznego lub wewnętrznego (podsystemu homeostazy), na podstawie których tworzone są reprezentacje obserwacji (o) oraz nagród ($r_ś$ i r_h);

- $[o\ldots]_{t+1}$ – zbiór obserwacji ‘o’ odnoszących się do bieżącego stanu środowiska, utworzony na podstawie bodźców $b_{ś\ t+1}$ oraz $b_{h\ t+1}$; poszczególne obserwacje przekazywane są do podsystemu monitorowania i motywacji, którego głównym zadaniem jest rozpoznawanie obserwacji relewantnych z perspektywy realizowanego celu oraz ignorowanie obserwacji nieistotnych; dlatego po przejściu przez podsystem P-MM zbiór [o…] redukowany jest symbolicznie do pojedynczej obserwacji ‘o’, istotnej z perspektywy celu; tego typu obserwacja, podobnie jak w wersji 1.0 modelu, umożliwia podsystemowi P-RL utworzenie wynikającej z niej reprezentacji stanu środowiska ‘s’;
- $[r_{ś}\ldots]_{t+1}$ – zbiór nagród reprezentujący natychmiastową zwrotną informację wartościującą na temat bieżącego stanu środowiska; wartość nagrody wyznaczana jest przez moduł ewaluacji zawarty w podsystemie P-SE, który „wycenia” napływające informacje, uwzględniając przy tym dane pochodzące z podsystemu homeostazy, tzn. odpowiednio zwiększa lub zmniejsza wartość nagrody $r_{ś}$ w zależności od tego czy organizm jest w stanie równowagi, czy jest zaburzony; podsystem P-MM „filtruje” dostępne nagrody, podobnie jak w przypadku zbioru [o…], udostępniając podsystemowi uczenia się ze wzmacnianiem wyłącznie nagrodę, która jest relewantna z perspektywy realizowanego celu;
- $[r_h\ldots]_{t+1}$ – nagroda ‘r_h’ jest utworzona na podstawie bodźców pochodzących z podsystemu homeostazy, reprezentuje natychmiastową informację zwrotną wartościującą, która odnosi się do bieżącego stanu organizmu; ten typ nagrody pozwala organizmowi realizować cele związane z zabezpieczeniem jego podstawowych potrzeb, w tym m.in. potrzebę bezpieczeństwa, bliskości, itp.; ponadto, ten typ nagród sygnalizuje podsystemowi monitorowania i motywacji przypadki naruszenia stanu homeostazy – np. braki energetyczne organizmu powodują pojawienie się stanu głodu odczuwanego jako nieprzyjemny;
- δ_{t+1} – błąd predykcji nagrody dla stanu s_{t+1} jest obliczany w ramach podsystemu uczenia się ze wzmacnianiem; służy do optymalizacji strategii doboru zachowań oraz informowania podsystemu monitorowania i motywacji o ewentualnych niedoszacowaniach lub przeszacowaniach danego stanu świata w odniesieniu do realizowanego celu ‘c’;
- c_x / c_y – operacja, która polega na dezaktywacji celu ‘x’ oraz na aktywacji celu ‘y’ w podsystemie P-RL; inicjatorem tego typu operacji jest podsystem monitorowania

i motywacji, który na podstawie asocjacji typu „obserwacja-nagroda-cel" ([o,r,c]) decyduje o tym, kiedy – na podstawie informacji wartościującej 'r' lub obserwacji 'o' – należy aktywować cel 'c'.

Symbole dodane do Modelu 1.2:

- O_{t+1} – reprezentacja stanu środowiska w chwili t+1, utworzona przez podsystem sensoryczno-ewaluacyjny dysponujący zdolnością uczenia się nowych typów reprezentacji; obserwacja typu 'O' – w odróżnieniu od obserwacji wrodzonych 'o' – ma charakter dynamiczny i zmienia się w czasie – wraz z dojrzewaniem organizmu;
- A ([o/O, $r_ś/r_h$/R → c], [o/O, R → Z]) – asocjacje łączące obserwacje, które dotyczą środowiska z nagrodami oraz celami (szczególnym przypadkiem celu może być wysokopoziomowe zachowanie Z – więcej na ten temat w dalszej części rozdziału); w podsystemie P-MM asocjacje umożliwiają: (1) aktywowanie celów 'c' na podstawie określonych obserwacji 'o/O' oraz związanych z nimi nagród lub (2) stabilizowanie celów poprzez filtrację obserwacji i nagród nieistotnych z perspektywy ich realizacji;
- Z – reprezentacja zachowania wyższego poziomu (patrz: koncepcja opcji zaprezentowana w rozdziale drugim w sekcji: *Hierarchiczne uczenie się ze wzmacnianiem*) utworzona przez podsystem zarządzający zachowaniami wyższego poziomu (P-ZZWP) w trakcie rozwoju ontogenetycznego; niskopoziomowe reprezentacje zachowań (z_t), skorelowane z nimi reprezentacje stanów świata (s_t) oraz obserwacje (O/o_t) – wraz z relacjami istniejącymi między nimi (patrz: korelacja „s_t, z_t → s_{t+1}, δ_{t+1}") – umożliwiają utworzenie tego typu reprezentacji (Z);
- R_Z – reprezentacja określająca wartość zachowania wyższego poziomu (Z), wyznaczona przez moduł ewaluujący podsystemu P-SERU; za pomocą tego typu reprezentacji zachowanie Z zaczyna być traktowane jak nagroda, która wpływa na dobór zachowań w podsystemie P-H-RL.

Symbole dodane do Modelu 2.0:

- SI – reprezentacja stanu intencjonalnego utworzona w ramach podsystemu zarządzania siecią stanów intencjonalnych; stan intencjonalny w ujęciu Searle'a posiada strukturę, która decyduje o tym, w jaki sposób odnosi się on do rzeczywistości i jaką treść zawiera; równocześnie poszczególne stany wzajemnie się warunkują i tworzą ze sobą sieć relacji;
- R_{SI} – reprezentacja wartości nagrody związanej z danym stanem intencjonalnym SI; wartość R_{SI} wyznaczana jest – podobnie jak w poprzednich modelach – przez P-SERU, zgodnie z hipotezą wspólnej neuronalnej waluty.

Symbole dodane do Modelu 3.0:

- Z – zachowanie wysokiego poziomu utworzone przez podsystem P-PRP (tzw. opcja);
- r_k – nagrody kształtujące służące do przekazywania wiedzy domenowej do podsystemu P-H-RL-OD, istotnej z perspektywy realizowanego planu.