WYJAŚNIENIE ZA POMOCĄ REPREZENTACJI MENTALNYCH

MONOGRAFIE
FUNDACJI NA RZECZ NAUKI POLSKIEJ

FUNDACJA NA RZECZ NAUKI POLSKIEJ

Paweł Gładziejewski

WYJAŚNIENIE ZA POMOCĄ REPREZENTACJI MENTALNYCH

PERSPEKTYWA MECHANISTYCZNA

WARSZAWA–TORUŃ 2015

Książka wydana przez
Fundację na rzecz Nauki Polskiej
w ramach programu Monografie FNP

Redaktor tomu
Anna Mądry

Korekty
Magdalena Bizior-Dombrowska

Projekt okładki i obwoluty
Barbara Kaczmarek

ISBN 978-83-231-xxxx-x

WYDAWNICTWO NAUKOWE
UNIWERSYTETU MIKOŁAJA KOPERNIKA
Redakcja: ul. Gagarina 5, 87-100 Toruń
tel. +48 56 611 42 95, fax +48 56 611 47 05
e-mail: wydawnictwo@umk.pl
Dystrybucja: ul. Reja 25, 87-100 Toruń
tel./fax: +48 56 611 42 38, e-mail: books@umk.pl
www.wydawnictwoumk.pl

Wydanie pierwsze
Druk: Wydawnictwo Naukowe UMK
ul. Gagarina 5, 87-100 Toruń

Moim rodzicom, Danucie i Grzegorzowi

Spis treści

PRZEDMOWA 11
WSTĘP 15
ROZDZIAŁ 1. REPREZENTACJE MENTALNE: KRAJOBRAZ PROBLEMÓW 25
1.1. Reprezentacjonizm i antyreprezentacjonizm w kognitywistyce 25
1.1.1. Reprezentacjonizm w kognitywistyce: krótka charakterystyka 25
1.1.2. Kognitywistyka antyreprezentacjonistyczna 29
1.1.3. Spór reprezentacjonizm–antyreprezentacjonizm: poziom przedmiotowy i metaprzedmiotowy 40
1.2. Spór o reprezentacjonizm a naturalizowanie intencjonalności 48
1.3. W poszukiwaniu koncepcji wyjaśniania reprezentacyjnego 58
ROZDZIAŁ 2. MECHANISTYCZNY MODEL WYJAŚNIANIA NAUKOWEGO 67
2.1. Mechanicyzm: ogólna charakterystyka 67
2.1.1. Wyjaśnianie mechanistyczne. Natura, badanie i naukowe reprezentowanie mechanizmów 67
2.1.2. Mechanicyzm, nomologiczno-dedukcyjny model wyjaśniania i pluralizm eksplanacyjny 84
2.2. Wyjaśnianie mechanistyczne a relacje międzypoziomowe 90
2.2.1. Poziomy w nauce jako poziomy mechanizmów 90
2.2.2. Relacje międzypoziomowe a granice mechanistycznej redukcji 96
2.3. Aplikacja modelu mechanistycznego do wyjaśniania w kognitywistyce 107
2.3.1. Wyjaśnianie mechanistyczne w kognitywistyce: zasadnicze założenia 107
2.3.2. Mechanicyzm, wyjaśnienia funkcjonalne i wieloraka realizowalność 111
ROZDZIAŁ 3. REPREZENTACJE I MECHANICYZM. PROBLEM MECHANIZMÓW REPREZENTACYJNYCH 121
3.1. Wyjaśnianie mechanistyczne za pomocą reprezentacji: uwagi wstępne 121

3.1.1.Wyjaśnianie reprezentacyjne jako wyjaśnianie za pomocą mechanizmów reprezentacyjnych 121
3.1.2. Mechanicyzm i problem reprezentacji: kilka kwestii dodatkowych 126
3.2. Funkcjonalne rozumienie reprezentacji i metoda Ramseya 134
3.2.1. Reprezentacje jako funkcjonalne komponenty mechanizmów. W poszukiwaniu funkcjonalnej koncepcji reprezentacji 134
3.2.2. Czym jest funkcja reprezentowania? Wymóg opisu zadań i metoda Ramseya 142
3.3. Metoda Ramseya: przykłady zastosowań negatywnych..... 161
3.3.1. Krytyka reprezentacji receptorowych 161
3.3.2. Krytyka reprezentacji ukrytych 172
ROZDZIAŁ 4. KONCEPCJA MECHANIZMÓW REPREZENTACYJNYCH 181
4.1. Od zewnętrznych reprezentacji ikonicznych do wewnętrznych S-reprezentacji 181
4.1.1.Reprezentacje w kognitywistyce a Peirce'owska triada 181
4.1.2. Reprezentacje wewnętrzne jako reprezentacje ikoniczne: ustalenia wstępne 190
4.1.3. Wewnętrzne reprezentacje ikoniczne: problem relacji podobieństwa. Pojęcie reprezentacji strukturalnych 196
4.1.4. Wewnętrzne reprezentacje ikoniczne i problem roli funkcjonalnej. S-reprezentacje bez interpretujących podmiotów 204
4.2. Mechanizmy reprezentacyjne jako mechanizmy wyposażone w konsumowane modele 215
4.2.1. Mechanizmy, S-reprezentacje i konsumowane modele 215
4.2.2. Teoria mechanizmów wykorzystujących konsumowane modele. Zarzuty i odpowiedzi 230
4.3. Mechanizmy reprezentacyjne: zastosowania w kognitywistyce 266
ROZDZIAŁ 5. REPREZENTACJONIZM W KOGNITYWISTYCE A PROBLEM NATURALIZACJI INTENCJONALNOŚCI 295
5.1. Naturalizowanie intencjonalności i dystynkcja osobowe––subosobowe 295
5.1.1. Dwa obrazy Sellarsa a osobowy i subosobowy poziom wyjaśniania 295
5.1.2. Naturalizowanie intencjonalności a założenie o korespondencji osobowe–subosobowe 303

5.1.3. Jak rozumieć relację między poziomem osobowym a subosobowym? O potrzebie alternatywy dla założenia korespondencji 316
5.2. Poziom osobowy i subosobowy: interpretacja mechanistyczna 335
5.2.1. Poziom osobowy i subosobowy jako poziomy mechanizmów 335
5.2.2. Interludium: McDowell o poziomie osobowym i wyjaśnieniach konstytutywnych 350
5.2.3. Konsekwencje mechanistycznego odczytania dystynkcji osobowe–subosobowe 355
5.3. Postawy propozycjonalne jako własności wyższego rzędu 368
5.3.1. Postawy propozycjonalne, zobowiązania architektoniczne i „miękka" naturalizacja. Ustalenia wstępne 368
5.3.2. Postawy propozycjonalne jako własności dyspozycyjne systemów poznawczych 379
5.3.3. Postawy propozycjonalne jako przyczyny 397
ZAKOŃCZENIE 417
SPIS RYSUNKÓW I TABEL 429
BIBLIOGRAFIA 431
SUMMARY 449

Przedmowa

Przedmiotem tej książki jest relacja umysłu do świata. Podejmuję w niej próbę rozwinięcia filozoficznej idei, która leży u podstaw całej nowożytnej teorii poznania i głosi, że nasz kontakt poznawczy ze światem nie jest bezpośredni, lecz zapośredniczony za pomocą wewnętrznych, umysłowych *reprezentacji* tego świata. Pragnę tu przekonać czytelnika do trzech twierdzeń. Po pierwsze, wspomniana idea może być w sposób precyzyjny wyrażona w kontekście kognitywistyki, czyli interdyscyplinarnej nauki, w której (między innymi) filozofowie, neuronaukowcy, psychologowie, specjaliści od sztucznej inteligencji oraz językoznawcy podejmują wspólny wysiłek, aby dostarczyć *stricte* naukowego (naturalistycznego) wyjaśnienia zdolności poznawczych i umysłowych. Dążę do pokazania, jak pojęcie reprezentacji mentalnych (umysłowych), mające przecież czysto filozoficzny rodowód, może odegrać rolę w wysiłkach badawczych kognitywistów. Po drugie, staram się pokazać, że – wbrew twierdzeniom bronionym przez współczesnych antyreprezentacjonistów – reprezentacje nie tylko mogą odgrywać rolę w kognitywistycznych wyjaśnieniach, ale także że taką rolę w istocie odgrywają. Po trzecie, próbuję bronić twierdzenia, że reprezentacje postulowane przez kognitywistów w zasadniczym stopniu różnią się od stanów, takich jak przekonania, pragnienia, intencje, oczekiwania czy wątpliwości – słowem, od stanów reprezentacyjnych, za pomocą których zwykli ludzie wyjaśniają własne oraz cudze działania w ramach codziennych interakcji społecznych. Jeśli mam rację, to zarówno istnienie stanów tego ostatniego rodzaju, jak i wartość wyjaśniania działań przez odwołanie się do nich, są w dużym stopniu niezależne od tego, co kognitywiści odkryją na poziomie wewnętrznych, neuronalnych czy neuroobliczeniowych mechanizmów poznawczych.

Pisząc tę książkę, starałem się pogodzić teoretyczne zaawansowanie bronionych twierdzeń z dostępnością mojego wywodu dla

niespecjalistów. Mam więc nadzieję, że choć ta publikacja jest pozbawiona jakichkolwiek uproszczeń, to będzie ona zrozumiała dla studentów kognitywistyki lub filozofii oraz dla pracowników akademickich – zarówno przedstawicieli humanistyki, jak i nauk szczegółowych – którzy co prawda filozofii lub kognitywistyki na co dzień nie uprawiają, jednak są zainteresowani problematyką sytuującą się na pograniczu tych dziedzin. Będzie mi niezwykle miło, jeśli ta książka okaże się przystępna i ciekawa przynajmniej dla części odbiorców, którzy nie parają się profesjonalnie nauką.

Przedstawione tu twierdzenia, argumenty oraz rekonstrukcje i analizy myśli innych autorów stanowią rezultat badań nad problemem reprezentacji, jakie podejmowałem w trakcie ostatnich kilku lat pod kierunkiem prof. Urszuli Żegleń w Zakładzie Kognitywistyki i Epistemologii Instytutu Filozofii UMK. Badania prowadziłem też w Stanach Zjednoczonych na Rutgers University, gdzie mogłem spędzić sześć miesięcy dzięki stypendium przyznanemu mi przez Polsko-Amerykańską Komisję Fulbrighta.

Zarówno na proces rozwijania bronionych tu pomysłów filozoficznych, jak i na proces pisania tej książki wpłynęło wiele osób, którym jestem niezmiernie wdzięczny. Szczególnie wiele zawdzięczam wspomnianej już prof. Żegleń. Nie tylko była ona promotorką mojej pracy doktorskiej, na której opiera się ta książka, ale wcześniej, jeszcze w trakcie studiów magisterskich, wprowadzała mnie w problematykę filozofii umysłu i kognitywistyki jako mój tutor w ramach Międzywydziałowych Indywidualnych Studiów Humanistycznych na UMK. Pani Profesor, dziękuję za naukę, pomoc i wsparcie przez te wszystkie lata.

Dziękuję też innym pracownikom Zakładu Kognitywistyki i Epistemologii IF UMK, dr. Tomaszowi Komendzińskiemu oraz dr Anicie Pacholik-Żuromskiej, za udział w seminariach doktorskich poświęconych dyskusji moich pomysłów. Doktorowi Komendzińskiemu zawdzięczam dodatkowo udostępnienie mi całego stosu anglojęzycznych książek, których lektura odcisnęła pozytywne piętno na prowadzonych przeze mnie badaniach. Chcę wyrazić wdzięczność (siłą rzeczy w języku, którego on nie zna) prof. Alvinowi Goldmanowi, który był dla mnie niezwykle pomocnym – przede wszystkim na-

ukowo, ale też pozanaukowo – opiekunem w trakcie mojego pobytu na Rutgers University. Dziękuję też prof. Marcinowi Miłkowskiemu za wyrażone przy różnych okazjach uwagi krytyczne do pewnych moich pomysłów filozoficznych, jak również za to, że tuż po tym, jak obroniłem doktorat, zaproponował mi odbycie pod jego naukowym okiem stażu podoktorskiego w Instytucie Filozofii i Socjologii PAN.

Osobne podziękowania składam Ani Karczmarczyk oraz (kiedy piszę te słowa, już doktorowi) Przemkowi Nowakowskiemu, moim przyjaciołom, z którymi przez lata dzieliłem doktorancki los. Jestem im wdzięczny za godziny dyskusji – choć trzeba uczciwie przyznać, często schodzących na tematy pozanaukowe – i za wsparcie w momentach zwątpienia.

Książka ta miała aż czterech recenzentów. Jej pierwotna wersja, która była moją pracą doktorską, została zrecenzowana przez prof. Roberta Piłata i prof. Roberta Poczobuta. Kolejna, poprawiona wersja została oceniona przez dwóch anonimowych recenzentów dla Rady Wydawniczej programu Monografie Fundacji na rzecz Nauki Polskiej. Każda z recenzji była zarazem życzliwa, jak również pełna niezwykle cennych uwag krytycznych, dzięki którym mogłem znacząco uzupełnić, poprawić i wzbogacić tę książkę. Wszystkim czterem recenzentom serdecznie dziękuję.

Chciałbym wreszcie podziękować mojej rodzinie. W trakcie pracy nad doktoratem będącym podstawą tej książki, moja narzeczona, Emilia, stała się moją żoną. Jestem jej wdzięczny za to, jak systematycznie wspierała mnie na duchu w trakcie pisania oraz za to, jak dzielnie znosiła fakt, że nawet kiedy nie pisałem, moje myśli zbyt często wędrowały w kierunku problemu reprezentacji mentalnych. Dziękuję moim rodzicom, Danucie i Grzegorzowi, za stworzenie mi warunków życia, w których mogłem swobodnie rozwijać swoje pasje i zainteresowania. Ostatecznie to im zawdzięczam fakt, że mogłem zająć się pracą naukową.

Wstęp

Jedną z ważnych idei sformułowanych przez Wilfrieda Sellarsa jest rozróżnienie na „naukowy" oraz „manifestujący się" obraz świata (Sellars 1963). Pierwszy z nich przedstawia „atomy w próżni": to wizja świata wyłaniająca się z najlepiej uzasadnionych teorii i modeli z zakresu nauk szczegółowych. Drugi natomiast to „zdroworozsądkowa" wizja zawarta w potocznym, codziennym oglądzie świata. Ten ostatni obraz przedstawia świat jako wypełniony obiektami, relacjami i własnościami, które na ogół nie widnieją w ramach obrazu naukowego. Obiekty codziennego doświadczenia są chociażby, w przeciwieństwie do obiektów, o których mówi fizyka, kolorowe. Co więcej, świat przedstawiony w obrazie manifestującym się jest także wypełniony bytami szczególnego rodzaju, dla których, jak się wydaje, nie ma miejsca w obrazie naukowym: świadomymi, myślącymi podmiotami, żyjącymi w świecie norm, a nie jedynie przyczyn. Według Sellarsa uzgodnienie tych dwóch fundamentalnie różnych i pozornie niekompatybilnych obrazów stanowi centralne wyzwanie filozofii. Filozofowie powinni dążyć do stworzenia synoptycznej wizji pokazującej, jak można utrzymywać realność świata zamieszkiwanego przez ludzi w ich codziennym doświadczeniu – świata kolorowych przedmiotów fizycznych, norm społecznych i podmiotów stanów mentalnych – pozostając jednocześnie realistą względem wizji zawartej w obrazie naukowym.

Jak zostało zaznaczone, jeden z elementów wchodzących w skład obrazu manifestującego się stanowi idea, że świat jest pełen specyficznych istot – posiadaczy umysłów. Obraz ten zawiera też określoną wizję tego, na czym polega posiadanie umysłu czy też bycie podmiotem stanów mentalnych. Wizja ta przedstawia ludzi jako istoty, których działania są kierowane stanami charakteryzującymi się, między innymi, *intencjonalnością*. Stany te – do których należą przekonania, pragnienia, intencje, oczekiwania oraz inne tak zwane posta-

wy propozycjonalne – posiadają treści czy też *reprezentują* określone stany rzeczy. Ludzie działają przecież na podstawie tego, że *sądzą*, iż rzeczy mają się tak, a nie inaczej, przy czym czasem sądzą oni niepoprawnie, czyli posługują się reprezentacjami błędnymi. Zrozumienie tego faktu nie wymaga formalnej edukacji psychologicznej, lecz stanowi fundament „zdroworozsądkowej", preteoretycznej aparatury pojęciowej, którą ludzie posługują się w celu regulowania swoich interakcji społecznych oraz przewidywania ich przebiegu.

Ów zestaw kategorii mentalnych (intencjonalnych) wykorzystywanych, aby nawigować własnym działaniem w ramach życia społecznego, określa się na ogół jako „psychologię potoczną". Wiele szczegółowych zagadnień dotyczących natury psychologii potocznej jest przedmiotem kontrowersji. Nie ma konsensu co do tego, czy zdolność posługiwania się „zdroworozsądkowymi" kategoriami psychologicznymi opiera się na znajomości *quasi*-naukowej – i w większości nieuświadamianej – teorii, czy też u jej podstaw stoi mechanizm innego typu, na przykład rodzaj symulacji mentalnej (por.: Goldman 2006; Carruthers 2009). Problematyczne pozostają też ewolucyjne źródła psychologii potocznej oraz kwestia tego, czy – i w jakim sensie – ma ona wrodzony charakter (por.: Gerrans 2002; Sterelny 2003: 211–240). Te szczegółowe zagadnienia są tu jednak nieistotne. Chodzi obecnie jedynie o zaznaczenie, że jednym z centralnych elementów manifestującego się obrazu świata jest koncepcja umysłu jako działającego na podstawie stanów intencjonalnych, czyli reprezentacji.

Jaką wizję umysłu przedstawia obraz naukowy? Czy istnieje w nim miejsce dla intencjonalności? Na pierwszy rzut oka wydawać się może, że tak. Pojęcie reprezentacji mentalnych przez dziesięciolecia stało przecież u podstaw naukowego rozumienia umysłu dostarczanego nam przez kognitywistykę[1]. Reprezentacje mentalne zdawały się pełnić rolę nie tylko w potocznych wyjaśnieniach ludzkich działań, ale też w rygorystycznie naukowych koncepcjach dotyczących tego, jak działa umysł (rozumiany jako system poznawczy,

[1] Terminy „kognitywistyka", „nauki kognitywne" oraz „nauki o poznaniu" są w tej książce stosowane zamiennie.

czyli system fizyczny zdolny do realizowania funkcji poznawczych). Czy oznacza to, że w przypadku intencjonalności Sellarsowskie uzgodnienie obrazu manifestującego się z obrazem naukowym otrzymaliśmy „za darmo" – dzięki temu, że wywiedzione z psychologii potocznej pojęcie reprezentacji okazało się skuteczne dla celów naukowych? Niekoniecznie.

Niektórzy filozofowie rzeczywiście zaangażowali się w próby pokazania, że – oraz w jaki sposób – stany intencjonalne psychologii potocznej mogą być identyczne z czysto neurobiologicznie lub obliczeniowo scharakteryzowanymi, wewnętrznymi stanami systemu poznawczego (por. m.in.: Millikan 1984; Dretske 1986, 1988; Fodor 1987). Pozycja ta pozwalała na uznanie przynajmniej niektórych naukowych wyjaśnień formułowanych przez kognitywistów za nieco doprecyzowane wersje wyjaśnień, które są formułowane na co dzień przez użytkowników psychologii potocznej. Jednak inni filozofowie podjęli próby pokazania, że najlepiej uzasadnione naukowe koncepcje umysłu bądź to w ogóle obywają się bez pojęcia reprezentacji, bądź to postulowane przez nie reprezentacje zasadniczo różnią się od stanów postulowanych w ramach psychologii potocznej (por.: Stich 1983; Ramsey, Stich, Garon 1990). Autorzy ci wyciągali na tej podstawie wniosek, że intencjonalne kategorie psychologii potocznej muszą zostać wykreślone przez kognitywistów z naukowego obrazu świata. Wedle tych filozofów obraz naukowy wymusza na nas odrzucenie istotnego fragmentu obrazu manifestującego się, a Sellarsowski „koncyliacyjny" projekt pogodzenia obydwu obrazów nie może się powieść – przynajmniej nie w tym konkretnym przypadku. Jeszcze inni filozofowie starali się wreszcie pokazać, że istnieje zasadniczy rozdźwięk pomiędzy „osobowymi" a „subosobowymi" wyjaśnieniami ludzkich działań, polegający między innymi na tym, że stany subosobowe – czyli te, do których odwołują się kognitywiści – w ogóle nie mogą charakteryzować się czymś takim, jak posiadanie treści intencjonalnej (por. Hornsby 2000). Autorzy broniący takiego stanowiska – stanowiący, jak się wydaje, mniejszość we współczesnej, na ogół naturalistycznie zorientowanej filozofii umysłu – nie wykorzystywali jednak rozbieżności między psychologią potoczną a kognitywistyką do eliminacji stanów postulowanych w ramach

tej pierwszej, lecz jedynie do stwierdzenia ich całkowitej wzajemnej autonomii. Z takiej perspektywy, intencjonalność wyznacza punkt, w którym obraz manifestujący się oraz obraz naukowy nie mogą się stykać. Są to w pewnym sensie obrazy różnych światów, które nie sposób zawrzeć w jednej, synoptycznej wizji postulowanej przez Sellarsa.

Co także warte zaznaczenia, w trakcie ostatnich lat pojęcie reprezentacji mentalnych stało się problematyczne dla samych kognitywistów. W naukach kognitywnych prężnie rozwijają się bowiem podejścia do modelowania i badania aktywności systemu poznawczego, które całkowicie obywają się bez postulowania wewnętrznych reprezentacji. Niewykluczone zatem, że reprezentacje wkrótce znikną z grona narzędzi eksplanacyjnych nauk kognitywnych, a tym samym przestaną stanowić część naukowego obrazu świata. To kolejny powód, by sądzić, że projekt Sellarsowskiego uzgodnienia tego, co naukowe, z tym, co się manifestuje, jest – przynajmniej kiedy mówimy o reprezentacjach mentalnych – zagrożony.

Powyższe uwagi mogą sugerować, że na styku filozofii z kognitywistyką istnieje jakiś jeden „problem reprezentacji". W istocie mamy tu jednak do czynienia z plątaniną różnych, choć wzajemnie powiązanych, zagadnień. Co kognitywiści mają na myśli, kiedy twierdzą, że wyjaśniają dane zjawisko za pomocą reprezentacji mentalnych? Czy reprezentacje te przypominają postawy propozycjonalne? Jak powinniśmy rozumieć naturę wyjaśnień reprezentacyjnych w kognitywistyce? Czy najlepsze teorie lub modele z zakresu nauk kognitywnych wykorzystują wyjaśnienia reprezentacyjne? Czy powinny one z takich wyjaśnień korzystać? Jaka jest relacja między reprezentacyjnymi wyjaśnieniami w kognitywistyce a tymi formułowanymi za pomocą psychologii potocznej? Czy metafizyczny status stanów intencjonalnych psychologii potocznej zależy od wizji umysłu, jaka będzie zawarta w kompletnej kognitywistyce? Wszystkie te pytania dotyczą oddzielnych zagadnień, jednak wszystkie w ten czy inny sposób łączą się z ogólnie rozumianym problemem relacji zachodzącej między pojęciem reprezentacji zawartym w obrazie naukowym a tym przedstawionym w obrazie manifestującym się.

Zasadniczym celem przedstawionych tu rozważań jest eksploracja zagadnień filozoficznych powstających na drodze między potocznym, manifestującym się ujęciem intencjonalności a zagadnieniem roli reprezentacji mentalnych w kognitywistycznych wyjaśnieniach zjawisk poznawczych. Jak postaram się pokazać, jest to droga prowadząca od przekonań i pragnień do wewnętrznych, zaimplementowanych w ośrodkowym układzie nerwowym modeli mentalnych.

To niezwykle ogólny zarys obszaru problemowego prezentowanej książki, warto więc go teraz doprecyzować. Pierwszym – a zarazem głównym – problemem, którym chcę się tu zająć, jest rola eksplanacyjna reprezentacji w naszym naukowym obrazie świata dostarczanym przez nauki kognitywne. Przedmiotem przedstawionych tu rozważań będzie natura wyjaśniania reprezentacyjnego w kognitywistyce, czyli problem tego, na czym polega wyjaśnianie zjawisk za pomocą reprezentacji mentalnych w ramach obrazu naukowego. Dokładniej, chcę zapytać o to, na czym polegają wyjaśnienia zasadnie czy pełnoprawnie – a nie jedynie pozornie – reprezentacyjne. Jak się okaże, zagadnienie to może zostać potraktowane jako równoważne pytaniu o to, na czym miałby polegać fakt, iż system poznawczy (biologiczny mózg) korzysta w swoim funkcjonowaniu z wewnętrznych reprezentacji.

Drugim poruszanym tu problemem będzie pytanie o to, czy kognitywistyka daje nam podstawy, by sądzić, że system poznawczy rzeczywiście korzysta z reprezentacji. Mówiąc precyzyjniej, pragnę podjąć próbę wykorzystania wypracowanej tu koncepcji wyjaśniania reprezentacyjnego, aby rozwiązać problem tego, czy – biorąc pod uwagę stan teoretyczny współczesnej kognitywistyki – system poznawczy może być *de facto* uznany za reprezentacyjny. Czy głosy mówiące o rychłym upadku reprezentacjonizmu w kognitywistyce są uzasadnione? Czy kognitywistyka rzeczywiście przechyla się w kierunku antyreprezentacjonizmu?

Nie chcę jednak zamykać prowadzonych tu rozważań w obrębie filozoficznej refleksji nad naturą obrazu naukowego. Moją intencją jest także przedstawienie pewnej propozycji dotyczącej możliwości uzgodnienia obu Sellarsowskich obrazów tego, jak działa umysł

(system poznawczy). Trzecim problemem teoretycznym tej pracy jest bowiem kwestia relacji zachodzącej między stanami intencjonalnymi psychologii potocznej (i wyjaśnieniami powołującymi się na te stany) a reprezentacjami postulowanymi przez kognitywistów (i wyjaśnieniami powołującymi się na te reprezentacje). Czy nasz manifestujący się obraz umysłu wymaga naukowej „legitymizacji", to jest pokazania, że przekonania, pragnienia i inne postawy propozycjonalne mogą stanowić część kognitywistycznych wyjaśnień zjawisk poznawczych? Jakie byłyby filozoficzne konsekwencje sytuacji, w której taka legitymizacja okazałaby się niemożliwa?

Podejmując wymienione problemy teoretyczne, osią swoich rozważań uczynię określoną koncepcję tego, na czym polega wyjaśnianie w kognitywistyce. Będę się tu mianowicie odwoływać do mechanistycznego modelu wyjaśniania. Model ten w trakcie ostatnich lat zyskał ogromne znaczenie w filozoficznych rozważaniach nad wyjaśnianiem naukowym, w tym wyjaśnianiem w naukach kognitywnych. Zawężając się jedynie do kognitywistyki, zgodnie z takim nowoczesnym „mechanicyzmem" (1) system poznawczy stanowi wielopoziomowy, hierarchicznie zorganizowany układ mechanizmów, natomiast (2) wyjaśnianie działania tego systemu polega na odkrywaniu mechanizmów z niższego poziomu organizacji, które odpowiadają za zjawiska z poziomu wyższego. Twierdzenia te będą stanowić teoretyczne tło, a zarazem spoiwo prowadzonych tu rozważań. Sądzę bowiem, że koncepcja mechanistyczna niesie istotne konsekwencje dla wszystkich podejmowanych tu zagadnień. Pozwala ona przede wszystkim pokierować rozważaniami nad naturą wyjaśniania reprezentacyjnego w kognitywistyce. Jednocześnie dostarczana przez mechanicyzm wizja relacji międzypoziomowych w kognitywistyce pozwala sformułować rozwiązanie problemu zależności między naukową wizją umysłu/poznania a wizją zawartą w psychologii potocznej.

Warto na tym etapie naszkicować strukturę książki oraz przedstawić jej zawartość. Rozdział 1 w znacznym stopniu zostanie poświęcony pogłębieniu i uszczegółowieniu wątków poruszonych już we wstępie. Jak zauważyłem wyżej, „problem reprezentacji" interesujący filozofów i kognitywistów to w istocie cały szereg zagadnień,

częściowo zazębiających się, a częściowo odrębnych i niezależnych. Celem rozdziału 1 będzie sformułowanie oraz możliwie precyzyjne rozróżnienie poszczególnych problemów związanych z pojęciem reprezentacji mentalnych, które są współcześnie dyskutowane na styku filozofii z kognitywistyką, a które, niestety, często nie są od siebie odróżniane w literaturze. W rozdziale tym odróżnię przede wszystkim przedmiotowy (czyli dotyczący przydatności reprezentacji w praktyce eksplanacyjnej kognitywistów) i metaprzedmiotowy (czyli dotyczący samej natury wyjaśniania reprezentacyjnego) problem statusu eksplanacyjnego reprezentacji w kognitywistyce. Podejmę także kwestię relacji między dyskusjami nad użytecznością eksplanacyjną w kognitywistyce a realizowanym przez niektórych filozofów umysłu projektem naturalizacji intencjonalności. Wszystko to posłuży z kolei dokładniejszemu wyznaczeniu celów tej książki, jak również wstępnemu sformułowaniu bronionych w niej tez.

Rozdział 2 zostanie poświęcony rozwinięciu zaznaczonej wyżej idei, zgodnie z którą wyjaśnianie naukowe – w tym wyjaśnianie w kognitywistyce – polega na wskazywaniu mechanizmów odpowiedzialnych za zjawiska. Przedstawię tam ogólne założenia mechanistycznego modelu wyjaśniania, a następnie skupię się na tych jego elementach, które odegrają ważną rolę w dalszej części książki. Szczególną uwagę poświęcę mechanistycznemu ujęciu relacji międzypoziomowych w nauce. Omówię konsekwencje mechanicyzmu dla zagadnienia natury redukcji, w tym zagadnienia granic wyjaśniania redukcyjnego. Przyjrzę się też potencjalnym problemom, na jakie mogą natrafiać próby aplikacji modelu mechanistycznego do wyjaśniania w kognitywistyce.

W rozdziale 3 podejmę kwestię konsekwencji, jakie model mechanistyczny niesie dla problemu wyjaśniania reprezentacyjnego w kognitywistyce. Dokładniej, zadam tam pytanie o to, na czym polegają *reprezentacyjne* wyjaśnienia mechanistyczne (mechanistyczne wyjaśnienia odwołujące się do reprezentacji). Zgodnie z przyjętym tu stanowiskiem mechanistyczne wyjaśnienie reprezentacyjne postuluje, że u podstaw wyjaśnianego zjawiska stoi mechanizm reprezentacyjny. Ten ostatni dysponuje z kolei komponentem lub grupą komponentów, których rola funkcjonalna w ramach mecha-

nizmu polega na reprezentowaniu czegoś. Ta ogólna idea zostanie w rozdziale 3 rozwinięta i sprecyzowana za pomocą zaproponowanego przez Williama Ramseya (2007) „wymogu opisu zadań" (*job description challenge*). Wymóg opisu zadań to filozoficzne/metodologiczne „narzędzie", dzięki któremu możliwe staje się określenie, czy struktury postulowane jako reprezentacje w danym wyjaśnieniu mają profil funkcjonalny, który pozwala *zasadnie* uznać je za reprezentacje. Powołując się na rozważania Ramseya, pokażę też, że przynajmniej niektóre pojęcia reprezentacji, jakimi posługują się przedstawiciele nauk kognitywnych, nie spełniają wymogu opisu zadań.

W rozdziale 4 podejmę próbę wykorzystania mechanicyzmu do sformułowania pozytywnej koncepcji wyjaśniania reprezentacyjnego w naukach kognitywnych. Zgodnie z przedstawioną w tym rozdziale koncepcją mechanizmy reprezentacyjne to mechanizmy wykorzystujące reprezentacje *strukturalne*, czyli oparte na podobieństwie strukturalnym. Ujmując tę ideę nieco inaczej, mechanizmy reprezentacyjne korzystają z wewnętrznych modeli. Modele te nie tylko są strukturalnie podobne do tego, co reprezentowane (modelowane), ale też są wykorzystywane *jako* modele przez inny komponent czy grupę komponentów mechanizmu. Wyjaśnić pewne zjawisko za pomocą reprezentacji to wyjaśnić je za pomocą mechanizmu wykorzystującego konsumowany model pewnej domeny. Tak właśnie brzmi proponowane tu rozwiązanie pierwszego, głównego problemu teoretycznego tej pracy. W rozdziale 4 zaproponuję jednak także rozwiązanie drugiego ze stawianych tu zagadnień. Postaram się pokazać, że wyjaśnianie przez mechanizmy reprezentacyjne – czyli korzystające z wewnętrznych modeli – stanowi istotny element *rzeczywistej* praktyki eksplanacyjnej kognitywistów. To znaczy, iż istnieje w kognitywistyce liczna grupa konceptualnie rozwiniętych i empirycznie ugruntowanych wyjaśnień różnych zjawisk poznawczych, które to wyjaśnienia powołują się na wewnętrzne modele jako eksplananse. Twierdzenia o marginalizacji czy wręcz upadku reprezentacjonizmu we współczesnej kognitywistyce są przedwczesne i nieuzasadnione.

Przedmiotem moich rozważań w rozdziale 5 będzie ostatni problem teoretyczny tej pracy, czyli kwestia relacji między naukowym a manifestującym się (potocznym) obrazem umysłu/systemu poznawczego. Postaram się tam wskazać konsekwencje, jakie przyjęcie perspektywy mechanistycznej niesie dla dyskutowanego w filozofii umysłu problemu naturalizacji intencjonalności. Będę bronił tezy, zgodnie z którą u podstaw programu naturalizacji stoi określone, na ogół przyjmowane *implicite* założenie dotyczące relacji między poziomami wyjaśniania zjawisk poznawczych czy umysłowych. Dokładniej, próbom znaturalizowania intencjonalności towarzyszy często presupozycja, że zachodzi odpowiedniość między poziomem *osobowym* – na którym wyjaśniamy działania podmiotów na podstawie żywionych przez nie przekonań, pragnień i innych postaw propozycjonalnych – a poziomem *subosobowym*, na którym opisujemy neuronalne/obliczeniowe mechanizmy stojące u podstaw określonych kompetencji poznawczych. W rozdziale 5 koncepcja wyjaśniania mechanistycznego zostanie wykorzystana do wykazania, że taka wizja relacji międzypoziomowych pozostaje nieuzasadniona. Będę tam postulować, iż psychologia potoczna jest *mechanistycznie neutralna*, to znaczy koncentruje się ona na interakcjach, w jakie system poznawczy jako całość wchodzi ze środowiskiem zewnętrznym. Wyjaśnianie działań za pomocą pojęć mentalnych składających się na psychologię potoczną nie niesie ze sobą zobowiązań teoretycznych dotyczących natury wewnętrznych mechanizmów poznania. W takiej perspektywie postawy propozycjonalne powinny być potraktowane jako własności wyższego rzędu, które nie sposób zidentyfikować z wewnętrznymi stanami czy strukturami systemu poznawczego. Odgrywają one w związku z tym inne role eksplanacyjne niż (subosobowe) reprezentacje postulowane przez kognitywistów. Problem naturalizacji intencjonalności jest zatem bardziej autonomiczny względem rozstrzygnięć w obrębie kognitywistyki, niż na ogół przyjmują filozofowie. Dotychczasowe niepowodzenia projektów zmierzających do naturalizacji intencjonalności biorą się z faktu, że u ich podstaw leży nieuzasadnione założenie dotyczące relacji międzypoziomowych, a nie z faktu, że w naukowym obrazie świata

nie ma miejsca dla przekonań, pragnień i innych (osobowych) stanów intencjonalnych.

W zakończeniu zostanie zawarte panoramiczne podsumowanie prowadzonych tu rozważań oraz bronionych tez. Zarysuję tam także dalsze perspektywy badawcze, jakie wyłaniają się na gruncie zaproponowanych w tej książce rozstrzygnięć.

Rozdział 1

Reprezentacje mentalne: krajobraz problemów

1.1. Reprezentacjonizm i antyreprezentacjonizm w kognitywistyce

1.1.1. Reprezentacjonizm w kognitywistyce: krótka charakterystyka

Przyglądając się niektórym opisom encykopedycznym kognitywistyki oraz wprowadzeniom do tej dziedziny, możemy dojść do wniosku, że związek między naukami kognitywnymi a pojęciem reprezentacji mentalnej ma niemal definicyjny charakter. W sporządzonym przez Paula Thagarda (2010) dla *Stanford Encyclopedia of Philosophy* haśle *Cognitive Science*, idea wyjaśniania zjawisk i kompetencji umysłowych przez postulowanie wewnętrznych reprezentacji jest nazywana „centralną hipotezą" nauk kognitywnych. W swoim artykule Thagard omawia różne podejścia do zagadnienia natury reprezentacji oraz tego, jak są one przetwarzane, jednak samo założenie o istnieniu i ważnej roli eksplanacyjnej reprezentacji uznaje on za niekwestionowany fundament kognitywistyki jako dziedziny naukowej. Z takiego punktu widzenia reprezentacje mogą mieć charakter symboliczny bądź obrazowy; mogą być węzłami w sieci koneksyjnej lub wzorami pobudzeń populacji neuronów; mieć naturę pojęciową lub obrazową; jakkolwiek byśmy ich jednak nie rozumieli, reprezentacje mentalne stanowią podstawowe narzędzie eksplanacyjne kognitywistyki.

Podejście prezentowane przez Thagarda nie jest rzecz jasna odosobnione. Barbara von Eckhardt (1993: 13–56) w swojej książce *What is Cognitive Science?* wychodzi od ogólnej definicji kognitywistyki jako interdyscyplinarnej nauki, której przedmiotem wyjaśniania są ludzkie zdolności poznawcze. Autorka ta zauważa jednak, że taka charakterystyka jest nazbyt liberalna i nie pokazuje, na czym polega *differentia specifica* stanowiąca o odrębnej tożsamości ko-

gnitywistyki. Wedle von Eckhardt kognitywistykę wyróżniają w ten sposób określone założenia teoretyczne i metodologiczne, a jednym z nich jest właśnie założenie o reprezentacyjnym charakterze systemu poznawczego. Ma ono dla tej autorki przede wszystkim charakter metafizyczny, dotyczy bowiem *natury* umysłu (systemu poznawczego) jako reprezentacyjnego oraz zachodzących w nim procesów jako będących operacjami przeprowadzanymi na wewnętrznych reprezentacjach. Według von Eckhardt to założenie bezpośrednio pociąga jednak za sobą kolejne – o charakterze *metodologicznym*, zgodnie z którym kognitywiści wyjaśniają interesujące ich zjawiska za pomocą wewnętrznych reprezentacji. Jeśli spojrzymy na zawartość omawianej pracy, okazuje się, że napisana przez von Eckhardt, niemal czterystustronicowa próba udzielenia odpowiedzi na pytanie, czym jest kognitywistyka, mogłaby niemal równie dobrze zostać potraktowana jako próba odpowiedzi na pytanie, czym są, i czym mogą być, reprezentacje mentalne.

Ogólną postawę teoretyczną prezentowaną przez Thagarda i von Eckhardt nazywa się zazwyczaj „reprezentacjonizmem". Stanowisko to może być jednak rozumiane na dwa sposoby – *deskryptywny* bądź *preskryptywny* – w zależności od tego, czy teza o roli reprezentacji mentalnych w kognitywistyce zostaje sformułowana w porządku *de facto*, czy też *de iure*.

Na poziomie deskryptywnym reprezentacjonizm to teza głosząca, że ogół lub zdecydowana większość teorii, wyjaśnień i hipotez formułowanych przez kognitywistów opiera się na (przyjmowanych *explicite* lub *implicite*) założeniach o (1) reprezentacyjnej naturze systemu poznawczego oraz (2) tym, że pojęcie reprezentacji mentalnych stanowi dla kognitywistyki narzędzie eksplanacyjne o fundamentalnym znaczeniu. Zauważmy, że bardziej ostrożne sformułowanie reprezentacjonizmu deskryptywnego[1] wydaje się prawdziwą tezą empiryczną dotyczącą rzeczywistej praktyki pojęciowej i badawczej przedstawicieli nauk kognitywnych. Kognitywistyka powstała, gdy rozwój informatyki, sztucznej inteligencji oraz nowe propozycje teo-

[1] Chodzi o sformułowanie mówiące o „zdecydowanej większości", a nie „ogóle" teorii, wyjaśnień i hipotez.

retyczne z zakresu psychologii pozwoliły (w naukowo rygorystyczny sposób) zapełnić czarną skrzynkę behawiorystów bogatym „życiem wewnętrznym" – życiem polegającym na przeprowadzaniu procesów obliczeniowych na wewnętrznych reprezentacjach środowiska zewnętrznego. Na przestrzeni ostatnich dziesięcioleci w naukach kognitywnych toczono spory wokół różnych szczegółowych problemów związanych z reprezentacjami. Dyskusje te dotyczyły, między innymi, obliczeniowej architektury kodującej reprezentacje (tego, czy są one kodowane symbolicznie, czy też subsymbolicznie, jako wagi połączeń między węzłami w sieciach koneksyjnych), jak również tego, czy mają one charakter analogowy, czy też propozycjonalny (por.: von Eckhardt 1993; Thagard 2010). Różne stanowiska dotyczące takich szczegółowych zagadnień można jednak postrzegać jako alternatywne ruchy w jednej, reprezentacjonistycznej grze. Spory te toczyły się w obrębie założenia o reprezentacyjnej naturze systemu poznawczego; dotyczyły problemu, czym są reprezentacje, jednak nie towarzyszyły im wątpliwości co do tego, czy one w ogóle istnieją i czy mają do spełnienia w kognitywistyce ważną rolę eksplanacyjną.

Z przyjmowanej przeze mnie perspektywy zdecydowanie bardziej interesujący jest reprezentacjonizm rozumiany jako stanowisko preskryptywne. W tym sformułowaniu reprezentacjonizm to teza, zgodnie z którą kognitywistyka *powinna* traktować reprezentacje mentalne jako zasadniczy i niezbywalny element swojego repozytorium eksplanacyjnego. Najlepszym bądź też jedynym możliwym sposobem osiągnięcia celu epistemicznego kognitywistyki – polegającego na sformułowaniu poprawnych naukowych wyjaśnień zjawisk umysłowych i poznawczych – pozostaje posługiwanie się w swoich wyjaśnieniach pojęciem reprezentacji mentalnych. Bez postulowania wewnętrznych reprezentacji nie zrozumiemy, jak możliwa jest percepcja głębi, percepcja kategorialna, kontrola motoryczna, podejmowanie decyzji i planowanie działań, formowanie pojęć, czytanie innych umysłów, wyobraźnia czy inne zdolności poznawcze stanowiące eksplananda dla teorii i modeli budowanych przez kognitywistów.

Reprezentacjonizm preskryptywny objawia się na ogół nie tyle w otwartych deklaracjach, co raczej *implicite* w praktyce teore-

tycznej i badawczej kognitywistów. Widać go chociażby w tendencji Thagarda i von Eckhardt do tego, by traktować posługiwanie się pojęciem reprezentacji mentalnych jako warunek *sine qua non* uprawiania kognitywistyki jako takiej. Autorzy ci w zasadzie nie pozostawiają miejsca dla kognitywistyki *antyreprezentacjonistycznej*, nawet jeśli Thagard (2010) stwierdza, że praca wykonywana przez kognitywistów „w większości", ale nie w całości opiera się na założeniu o istnieniu reprezentacji mentalnych. Inny warty wspomnienia przykład reprezentacjonizmu preskryptywnego, tym razem pochodzący z badań eksperymentalnych, przytacza William Ramsey (2007: 147–148). Omawia on badania Waltera Freemana i Christine Skardy (1990), którzy prezentowali badanym określone zapachy, jednocześnie rejestrując za pomocą EEG aktywność neuronalną obszarów zaangażowanych w przetwarzanie sygnałów zapachowych. Badacze ci szukali korelacji między wzorami pobudzeń neuronalnych a rodzajami bodźców węchowych, kierując się założeniem, że mierzone przez nich aktywacje w jakiś sposób reprezentują zapachy, na które są eksponowani uczestnicy eksperymentu. Przez lata ich badania były niekonkluzywne – uzyskane wyniki nie poddawały się interpretacji w świetle przyjętych założeń teoretycznych. Jak się okazało, było temu winne zasadnicze założenie, że funkcją rejestrowanej aktywności neuronalnej jest *reprezentowanie* bodźców. Dopiero porzucenie reprezentacjonistycznych presupozycji dotyczących zgromadzonych danych pozwoliło Freemanowi i Skardzie na wyjście z eksperymentalnego impasu. Jak się wydaje, to właśnie duch współczesnej kognitywistyki, patrzącej na mózg jako system, którego nadrzędną funkcją jest generowanie i używanie wewnętrznych „odzwierciedleń" zewnętrznych zjawisk, pokierował pierwotnie te badania w niepożądanym kierunku. Dopiero uwolnienie się od reprezentacjonizmu i spojrzenie na mózg jako na system samoorganizujący się oraz działający w dużym stopniu na podstawie wygenerowanych wewnętrznie sygnałów pozwoliło badaczom na poczynienie postępu. Przykład Freemana i Skardy stanowi dobrą ilustrację tego, jak realny jest wpływ reprezentacjonizmu preskryptywnego – przyjętego w postaci ukrytego, niewyrażanego wprost założenia – na badania empiryczne prowadzone wewnątrz kognitywistyki.

Naturalne wydaje się przy tym założenie, że reprezentacjonizm preskryptywny, jako stanowisko zalecające określoną praktykę eksplanacyjną, jest motywowany określonymi założeniami dotyczącymi *natury* systemu poznawczego. Mówiąc wprost, przekonanie, iż wyjaśnienia odwołujące się do reprezentacji są pożądane oraz skuteczne, wydaje się motywowane założeniem, iż reprezentacje są realnymi, przyczynowo efektywnymi wewnętrznymi strukturami lub stanami systemu poznawczego. To natura systemu poznawczego gwarantuje sukces określonym strategiom wyjaśniania tego systemu. Można uznać, że reprezentacjonizm preskryptywny jako stanowisko epistemiczne (dotyczące wyjaśniania w kognitywistyce) splata się z metafizycznym twierdzeniem dotyczącym tego, czym są i jak działają rzeczywiste systemy poznawcze. Anthony Chemero (2009: 67–83) twierdzi, że ten związek jest filozoficznie problematyczny. Według tego autora przydatność eksplanacyjna pojęcia reprezentacji to dość zawodny wskaźnik tego, czy procesy umysłowe rzeczywiście korzystają z reprezentacji. Jest tak, ponieważ możliwe okazuje się według niego formułowanie reprezentacyjnych wyjaśnień systemów, które reprezentacyjne nie są. Nie podzielam stanowiska Chemero. W następnych rozdziałach postaram się pokazać, że przyjmując określone, wiarygodne założenia dotyczące tego, czym jest wyjaśnianie w naukach kognitywnych, reprezentacjonizm preskryptywny (jako tezę o pożądanej strategii eksplanacyjnej nauk kognitywnych) oraz twierdzenie o reprezentacyjnej naturze systemu poznawczego należy uznać za dwie strony jednej monety: skuteczność eksplanacyjna reprezentacji mentalnych jest możliwa (w świetle tych założeń) dzięki temu, że te są realne i rzeczywiście odpowiadają za zjawiska, które za ich pomocą wyjaśniamy. Reprezentacjonizm preskryptywny to zatem zalecenie metodologiczne, któremu towarzyszy twierdzenie o reprezentacyjnej naturze systemu poznawczego.

1.1.2. Kognitywistyka antyreprezentacjonistyczna

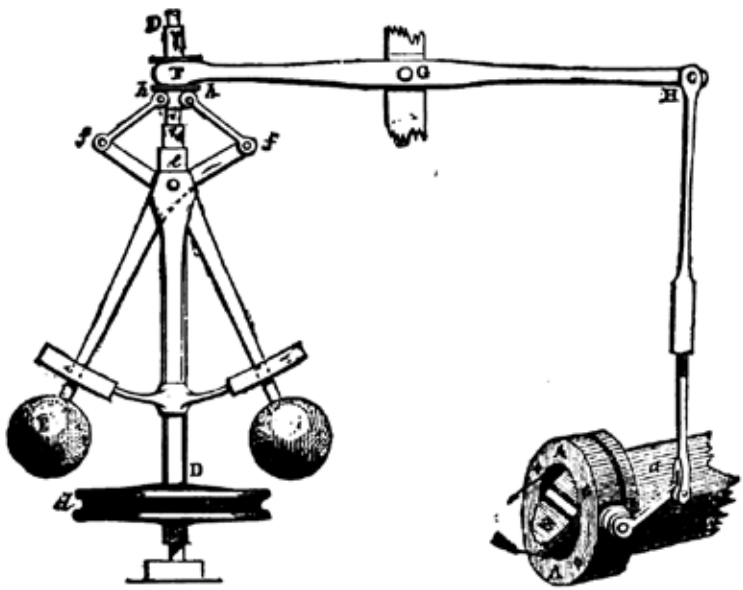
Jak zaznaczyłem, dokonane przez Thagarda i von Eckhardt zawężenie kognitywistyki do teorii i badań rozwijanych w duchu reprezentacjonizmu może zostać uznane za deskryptywnie poprawne

jedynie wtedy, gdy zinterpretujemy je ostrożnie, czyli jako tezę dotyczącą „większości", ale nie całości kognitywistyki. Chociaż główny nurt nauk kognitywnych, począwszy od lat pięćdziesiątych XX wieku, rzeczywiście rozwijał się pod „banderą" reprezentacjonizmu, to na jego marginesie pojawiały się także idee wyraźnie antyreprezentacjonistyczne. Nie wdając się w tej chwili w historyczne szczegóły, przyjrzyjmy się „programowemu" i klasycznemu już artykułowi, który odegrał znaczącą rolę w przybliżaniu podejścia antyreprezentacjonistycznego w kognitywistyce środowisku filozofów analitycznych. Mowa o opublikowanym w 1995 roku artykule *What might cognition be if not computation?* autorstwa Tima van Geldera[2].

Centralny punkt artykułu van Geldera (1995) stanowi dyskusja stosunkowo prostego układu mechanicznego, czyli rodzaju silnika parowego opracowanego w XVIII wieku przez Jamesa Watta. Silnik ten przekształcał stworzony pod wpływem ciśnienia ruch posuwisty tłoka znajdującego się w cylindrze w ruch obrotowy koła zamachowego. Dzięki temu silnik ten mógł bezpośrednio napędzać powszechne w tamtym czasie urządzenia sterowane za pomocą koła zamachowego. Inżynieryjne wyzwanie, przed jakim stanął Watt, polegało na wymyśleniu mechanizmu, dzięki któremu byłoby możliwe utrzymanie stałej prędkości generowanego przez silnik ruchu obrotowego. Co prawda cylinder silnika zawierał zawór regulujący dopływ pary do cylindra, a przez to umożliwiający pośrednio regulację prędkości koła zamachowego, jednak zawór ten nie działał automatycznie; musiał on być nadzorowany i obsługiwany przez człowieka. Szczególnie interesująca van Geldera innowacja dokonana przez Watta to właś-

[2] Tytuł tekstu van Geldera może sugerować, że głównym obszarem zainteresowania autora jest nie tyle problem reprezentacji, co raczej zagadnienie obliczeniowej natury procesów poznawczych czy umysłowych, a dokładniej: możliwość stworzenia nieobliczeniowej teorii tego, jak działa system poznawczy. W istocie, van Gelder traktuje reprezentacjonizm i komputacjonizm jako dwa elementy jednego „tradycyjnego" kognitywistycznego pakietu teoretycznego. (Warto nadmienić, że przekonanie, iż kognitywistyka jest/powinna być zarówno reprezentacjonistyczna, jak i komputacjonistyczna przyjmują we wspominanych już pracach Thagard i von Eckhardt). Dla bieżących celów analitycznych omawiany artykuł zostanie potraktowany przede wszystkim jako manifest antyreprezentacjonizmu, a jego antykomputacjonistyczny aspekt zostanie pominięty.

nie mechanizm pozwalający na czysto automatyczną – niewymagającą ludzkiej ingerencji – regulację dopływu pary do cylindra. Mowa o tak zwanym regulatorze Watta (*Watt governor*). Regulator ten stanowiło wrzeciono połączone z kołem zamachowym i cylindrem (por. rysunek 1). Do wrzeciona były podłączone dwa ramiona zakończone metalowymi kulami w taki sposób, że unosiły się one (a zatem ich kąt rozstawu powiększał się) bądź opadały (kąt pomniejszał się) w zależności od zmian w prędkości ruchu wrzeciona. Ramiona te z kolei były podłączone do zaworu regulującego dopływ pary do cylindra, w którym znajdował się tłok. Urządzenie działało, tak że wzrost prędkości ruchu obrotowego koła zamachowego prowadził do sytuacji, w której ramiona regulatora unosiły się, przymykając w ten sposób zawór doprowadzający parę do cylindra. W ten sposób prędkość ruchu koła zamachowego zmniejszała się. W sytuacji, gdy ustawał ruch koła, kąt rozstawu ramion regulatora stopniowo się zmniejszał, otwierając tym



Rysunek 1. Skonstruowany przez Jamesa Watta regulator kontrolujący działanie silnika parowego. Źródło: van Gelder 1992: 349

samym zawór i w efekcie prowadząc do sytuacji, w której prędkość ruchu koła zamachowego na nowo się zwiększała. Regulator Watta naprzemiennie otwierał i przymykał więc zawór doprowadzający parę do cylindra, zapewniając silnikowi w pełni zautomatyzowane, płynne działanie, w którym stale utrzymywano pożądaną prędkość generowanego ruchu obrotowego.

Jakie konsekwencje może mieć rozpatrzenie tak prostego układu mechanicznego dla zagadnienia statusu reprezentacji w kognitywistyce? Według van Geldera (1995) analiza działania silnika Watta[3] może przybliżyć nam sposób myślenia o systemie poznawczym, który radykalnie zrywa z wieloma zasadniczymi założeniami teoretycznymi i metodologicznymi „klasycznej" kognitywistyki. Van Gelder zauważa najpierw, że gdybyśmy chcieli wyjaśnić działanie silnika Watta za pomocą strategii eksplanacyjnej dominującej w kognitywistyce[4], zapewne uznalibyśmy, że regulator utrzymuje stałą prędkość ruchu generowanego przez silnik dzięki obliczeniowym operacjom prowadzonym na wewnętrznych, symbolicznych reprezentacjach. Proces kontroli zaczynałby się od pomiaru aktualnej prędkości ruchu koła zamachowego i porównania jej z prędkością pożądaną. Gdyby te wartości się różniły, dochodziłoby do pomiaru aktualnego ciśnienia w tłoku i oszacowania, na podstawie określonych algorytmów, ciśnienia pożądanego. Ostatecznie, zostałaby obliczona pożądana zmiana stopnia, w jakim jest otwarty zawór regulujący dopływ

---

[3] Za pomocą terminu „silnik Watta" będę tu skrótowo oznaczać silnik parowy wyposażony w regulator Watta.

[4] Pamiętajmy, że omawiany artykuł pochodzi z 1995 roku. W chwili, gdy van Gelder pisał swój tekst, dominujące w kognitywistyce podejście do wyjaśniania różnych zjawisk poznawczych nadal było wyznaczane przez ideę, że procesy poznawcze to formalne, obliczeniowe operacje na czysto syntaktycznie scharakteryzowanych symbolach. To właśnie jest „dominująca" strategia eksplanacyjna, o której mowa. Mimo to proponowana przez van Geldera idea silnika Watta jako modelu systemu poznawczego ma stanowić (i na ogół jest tak interpretowana) alternatywę dla dowolnej postaci reprezentacjonizmu. Tym samym artykuł van Geldera ma uderzać we wszelkich zwolenników reprezentacjonizmu, także tych, którzy nie są zwolennikami klasycznej, symboliczno-obliczeniowej koncepcji tego, jak działa system poznawczy. W taki właśnie sposób artykuł van Geldera jest interpretowany w ramach tej pracy.

pary do cylindra. Po wprowadzeniu określonych zmian w zaworze proces zaczynałby się na nowo od pomiaru aktualnej prędkości ruchu koła zamachowego.

Nie ulega wątpliwości, że takie wyjaśnienie nie odpowiadałoby temu, w jaki sposób naprawdę działa silnik Watta. Mówiąc wprost, mechanizm działania tego silnika nie korzysta z niczego, co przypominałoby algorytmiczne operacje na symbolicznych reprezentacjach. Według van Geldera (1995) nie istnieje żaden eksplanacyjnie wartościowy sens terminu „reprezentacja", zgodnie z którym moglibyśmy powiedzieć, że silnik Watta jest systemem sterowanym czy kontrolowanym za pomocą reprezentacji. Autor ten podaje kilka racji za tym stanowiskiem. Po pierwsze, nieformalny i ogólny opis silnika pozwala nam według van Geldera na pełne zrozumienie działania tego systemu bez konieczności posługiwania się pojęciem reprezentacji. Jest to przypadek, w którym możemy zrozumieć interesujący nas system bez konieczności uznania go za system reprezentacyjny. Powołanie się na reprezentacje nie przyniesie nam po prostu żadnego zysku poznawczego w próbach zrozumienia, jak działa silnik Watta. Po drugie, moglibyśmy być może przyjąć czysto kowariancyjne[5] rozumienie reprezentacji i uznać, że współzmienność zachodząca między kątem rozstawu ramion regulatora a prędkością koła zamachowego sprawia, że ten pierwszy reprezentuje tę drugą. Jednak takie stanowisko okazuje się zbyt liberalne; istnieje wiele współzmienności, które nie są reprezentacjami, a zatem współzmienność nie wystarcza dla uczynienia czegoś reprezentacją. Po trzecie, gdybyśmy nawet przyjęli możliwie najbardziej liberalne, oparte na współzmienności pojęcie reprezentacji, to i tak nie pozwoli nam ono na uznanie silnika Watta za system posługujący się reprezentacjami. Jak bowiem zauważa van Gelder, między kątem rozstawu ramion regulatora a prędkością koła zamachowego w silniku Watta po prostu nie zachodzi współzmienność (przynajmniej nie w odpowiednim stopniu). Kiedy prędkość koła raptownie się zmniejsza, potrzeba nieco

[5] Mówiąc najprościej, chodzi o takie ujęcie, w którym *X* reprezentuje *Y*, jeśli *X* regularnie współzmienia się z *Y*. Jest to „receptorowe" rozumienie reprezentacji, które zostanie krytycznie omówione w rozdziale 3 (podrozdział 3.3).

czasu, aby kąt rozstawu ramion dopasował się do tej zmiany. Zanim do tego dojdzie, obydwie wartości są „rozregulowane". Dlatego też jedna z nich nie może reprezentować drugiej. Po czwarte wreszcie, regulator Watta nie pełni wedle van Geldera w silniku funkcji reprezentacji. Pojęcie reprezentacji bardziej zaciemnia, niż rozjaśnia relację między kątem rozstawu ramion regulatora a prędkością koła zamachowego. Kąt nachylenia ramion i prędkość koła zamachowego wpływają na siebie bezpośrednio i w sposób stały, tak że można powiedzieć, iż jedno zarówno determinuje, jak i jest determinowane przez drugie. Żeby opisać ich wzajemną współzależność, potrzebujemy wedle van Geldera alternatywnego, bardziej subtelnego i złożonego niż reprezentacje mentalne narzędzia pojęciowego.

Van Gelder (1995) postuluje, że takie lepsze narzędzie uzyskamy wtedy, gdy wykorzystamy strategię eksplanacyjną zasadniczo różniącą się od tej dominującej w „klasycznej" kognitywistyce, a wykorzystującą matematyczną teorię systemów dynamicznych. Strategia ta opiera się na potraktowaniu silnika Watta jako fizycznej realizacji opisanego matematycznie systemu dynamicznego. Oznacza to w praktyce, że powinniśmy scharakteryzować silnik Watta za pomocą szeregu zmiennych opisujących różne jego elementy czy aspekty działania (na przykład zmianę kąta nachylenia ramion regulatora czy prędkość poruszania się koła zamachowego). Podejście dynamiczne opiera się na poszukiwaniu równań, które opisywałyby sposób, w jaki zmienne, za pomocą których opisujemy system – w tym przypadku silnik Watta – są od siebie wzajemnie zależne. Równania te możemy potraktować jako dynamiczne „prawa" opisujące działanie interesującego nas systemu. Van Gelder ilustruje to równaniem określającym, w jaki sposób zmiana kąta rozstawu ramion regulatora Watta stanowi funkcję aktualnego kąta rozstawu, dotychczasowego sposobu zmieniania się tego kąta oraz aktualnej prędkości koła zamachowego. Ostatecznie działanie silnika Watta można przedstawić jako trajektorię w przestrzeni stanów, czyli przestrzeni, której poszczególne wymiary są wyznaczane przez możliwe wartości poszczególnych opisujących silnik zmiennych. Każdy punkt w przestrzeni stanów wyznacza jeden możliwy całościowy stan silnika, a trajektoria systemu w obrębie tej przestrzeni pozwala nam obra-

zować ewolucję jego zachowania w czasie. Naturalne i sztuczne systemy, które opisujemy w ten sposób, będą miały często tendencję do „wpadania" w określone, charakterystyczne trajektorie w tego rodzaju przestrzeni, nazywane „atraktorami". W przypadku silnika Watta atraktorem będzie stan równowagi, w którym jest utrzymywana stała, pożądana prędkość koła zamachowego.

Van Gelder (1995) stara się pokazać w swoim artykule, że opisane wyżej podejście pozwala nam uchwycić „dynamikę" działania silnika Watta w sposób deskryptywnie adekwatny i kompletny. Silnik ten okazuje się według niego klarownym przykładem płynnie działającego, samoregulującego się systemu, który nie potrzebuje reprezentacji do działania i ich nie wykorzystuje. Teoria systemów dynamicznych stanowi zaś narzędzie, za pomocą którego możemy uzyskać matematycznie precyzyjne wyjaśnienie działania silnika obywające się bez pojęcia reprezentacji.

Centralna teza artykułu van Geldera brzmi zatem: realne systemy poznawcze przypominają bardziej silnik Watta niż maszynę Turinga. Systemy poznawcze są systemami dynamicznymi, których zachowanie można opisać jako trajektorię w przestrzeni stanów, czyli tak samo jak zachowanie silnika Watta[6]. Dla van Geldera oznacza to także, że charakterystyka systemu poznawczego jako systemu dynamicznego nie będzie się odnosiła do niczego, co odgrywałoby rolę wewnętrznych reprezentacji. Ewolucja systemu poznawczego zostanie opisana jako trajektoria w przestrzeni stanów, a zrozumienie praw dynamicznych „rządzących" tą ewolucją obędzie się całkowicie bez pojęcia reprezentacji. Należy dodać, że opis taki będzie obejmował również relacje, w jakie badany system wchodzi ze środowiskiem zewnętrznym. Konkretnie wśród szeregu zmiennych potrzebnych do opisu systemu poznawczego będą takie, które odnoszą się do aspektów środowiska zewnętrznego i z którymi na różne

[6] Co oczywiste, systemy poznawcze są nieporównywalnie bardziej złożone niż silnik Watta i wymagają bardziej złożonego opisu. Mimo tej różnicy, w obu przypadkach identyczny ma być jednak rodzaj strategii, za pomocą której opisujemy i wyjaśniamy działanie badanego systemu.

sposoby powiązane (sprzężone) są zmienne opisujące sam system[7]. Co jednak istotne, taka charakterystyka relacji systemu poznawczego ze środowiskiem jest zasadniczo różna od tej, zgodnie z którą ten system *reprezentuje* swoje środowisko. W perspektywie dynamicznej – jak ją interpretuje van Gelder – system poznawczy nawiązuje ciągłe interakcje ze swoim środowiskiem, jednak te interakcje są bezpośrednie i nie wykorzystują żadnych poznawczych „pośredników" w postaci wewnętrznych reprezentacji.

Zanim będzie można przejść do dalszej, bardziej ogólnej charakterystyki antyreprezentacjonizmu, należy poczynić dwie istotne uwagi. Po pierwsze, omawiany tu artykuł van Geldera stanowi jedynie programowy zarys, mający przedstawić pewne nieklasyczne i niekonwencjonalne na tle kognitywistycznej „tradycji" spojrzenie na to, jak można opisywać i wyjaśniać działanie systemu poznawczego. Od czasu publikacji tego artykułu różni filozofowie próbowali rozwijać dynamiczne podejście od strony konceptualnej (por. chociażby: Calvo 2008; Chemero 2009, 2014). Próby te koncentrowały się często właśnie na pokazywaniu, w jakim dokładnie stopniu i w jakim sensie podejście dynamiczne jest wyrazem antyreprezentacjonizmu. Co więcej, dynamiczne podejście do badania procesów poznawczych okazało się płodne empirycznie. Istnieją interesujące badania pokazujące, że traktowanie systemów poznawczych jako systemów dynamicznych pozwala owocnie modelować proces podejmowania prostych praktycznych decyzji (Busemeyer, Townsend 1993), nabywanie kompetencji motorycznych (Thelen, Smith 1994) czy oszacowywanie postrzegalnych własności obiektów oraz przewidywanie skutków możliwych działań (van Rooij, Bongers, Haselager 2002). Niektóre z tych badań pokazują, że teoria systemów dynamicznych może mieć zastosowanie przy wyjaśnianiu zachowań zorientowanych na obiekty czy zdarzenia, z którymi osoba badana nie znajduje się aktualnie w bezpośrednim przyczynowym kontakcie (por. omówienie tego zagadnienia w: Chemero 2009: 47–66). Na przykład autorzy jednego ze wspomnianych już badań (van Rooij,

[7] W istocie system jest tak ściśle sprzężony ze środowiskiem, że nie sposób wyobrazić sobie, jak mógłby on funkcjonować poza nim.

Bongers, Haselager 2002) pokazali, że model dynamiczny może zostać z powodzeniem wykorzystany, aby przewidzieć decyzje osób badanych co do tego, czy za pomocą prezentowanego im patyka o określonej długości uda im się sięgnąć po pewien oddalony obiekt. Przedmiotem tego badania były zatem oceny badanych dotyczące jedynie możliwego, a nie realnie wykonywanego działania. Wyniki te są istotne dlatego, że mogą stanowić odpowiedź na głosy krytyków „dynamicznej" wersji antyreprezentacjonizmu. Krytycy ci sugerowali bowiem (Clark, Toribio 1994), że podejście dynamiczne nie może być stosowane do badania i modelowania procesów poznawczych uczestniczących przy realizacji zadań „złaknionych reprezentacji" (*representation-hungry*), czyli między innymi właśnie takich, które wymagają dostosowania działania do obiektów, zdarzeń czy procesów, które są przestrzennie i/lub temporalnie odległe na tyle, że nie mogą wpływać bezpośrednio na aktywność systemu poznawczego.

Po drugie, należy bardzo wyraźnie zaznaczyć, że w przedstawionej tu rekonstrukcji nie chodzi o proste utożsamienie antyreprezentacjonizmu z podejściem odwołującym się do teorii systemów dynamicznych. Antyreprezentacjonizm nie zawęża się do podejścia dynamicznego, jest kategorią szerszą. W historii nauk kognitywnych podejmowano próby rozwijania antyreprezentacjonistycznych koncepcji poznania, które były bardziej lub mniej odrębne w stosunku do podejścia dynamicznego. Należy tu z pewnością wspomnieć teorię percepcji bezpośredniej, sformułowaną przez Jamesa Gibsona, a opartą na idei, że organizmy mają bezpośredni, niemediowany reprezentacjami dostęp do potrzebnych im informacji o swoim środowisku (Gibson 1979). Kolejny często wymieniany w literaturze przykład antyreprezentacjonizmu to prace Randalla Beera (por. np.: 1997; 2003), pokazujące, jak wirtualne (symulowane) organizmy mogą inteligentnie działać w swoim środowisku bez konieczności wykorzystywania jego wewnętrznych reprezentacji. Za przykład podejścia antyreprezentacjonistycznego w robotyce poznawczej należy uznać z kolei roboty konstruowane przez Rodneya Brooksa, dla których świat nie musi być w żaden sposób wewnętrznie odzwierciedlany, bo stanowi on „swój własny model" (por. np. Brooks 1991). W literaturze filozoficznej historycznie ważne są z kolei prace Huberta

Dreyfusa (1972), który krytykował reprezentacjonizm w kognitywistyce z perspektywy teorii Maurice'a Merleau-Ponty'ego i Martina Heideggera, próbując na tej podstawie sformułować alternatywny, obywający się bez pojęcia reprezentacji mentalnych, model relacji między podmiotem (systemem poznawczym) a światem. Wszystkie wymienione koncepcje zostały sformułowane za pomocą różnych aparatów pojęciowych i w ramach bardzo różnych podejść do badania i opisywania poznania, jednak łączy je niechęć do reprezentacji jako narzędzia eksplanacyjnego. Co prawda niektóre z tych przykładów antyreprezentacjonizmu wykazują bardziej lub mniej wyraźne związki z podejściem dynamicznym[8]. Mimo to można stwierdzić, iż wcale nie jest tak, że „nie ma antyreprezentacjonizmu bez dynamizmu". Trzeba więc traktować perspektywę dynamiczną wyrażoną przez van Geldera jako jedną z potencjalnie wielu form, jakie może przybrać antyreprezentacjonizm w kognitywistyce.

Antyreprezentacjonistyczne idee na temat tego, jak powinniśmy rozumieć naturę poznania, długo znajdowały się na marginesie teoretycznej i badawczej praktyki kognitywistów. Trend ten jednak zaczął wyraźnie zmieniać się, począwszy od lat dziewięćdziesiątych XX, a zwłaszcza w pierwszej dekadzie XXI wieku. W tym okresie rosnącą popularność zaczęły zdobywać zdecydowanie bardziej „przyjazne" antyreprezentacjonizmowi sposoby rozumienia tego, jak działają systemy poznawcze. Zwolennicy koncepcji poznania ucieleśnionego podkreślają, że poznanie należy rozpatrywać w ścisłym powiązaniu z ciałem o określonych morfologicznych i motorycznych właściwościach (por.: Thompson, Rosch 1992; Gallagher 2005; Shapiro 2010; Wilson, Foglia 2011). Przedstawiciele usytuowanych lub zakorzenionych podejść w kognitywistyce postulują, że poznanie jest wynikiem ścisłych interakcji systemu poznawczego ze środowiskiem, zarówno fizycznym, jak i społeczno-kulturowym (por.: Kirsh, Maglio 1994; Hutchins 1995; Clark, Chalmers 2008; Robbins, Aydede 2008). Zwolennicy enaktywizmu twierdzą, że poznanie na-

[8] Najbardziej bezpośredni związek zachodzi w przypadku prac Beera (por. np.: 2003), który do opisu stworzonych przez siebie wirtualnych organizmów stosuje właśnie teorię systemów dynamicznych. Warte wspomnienia są też próby łączenia podejścia dynamicznego z teorią Gibsona (por. np.: Chemero 2009).

leży rozumieć jako formę aktywnego, eksploracyjnego działania, mającego charakter w jakimś sensie „konstruktywny", a nie będącego jedynie pasywnym rejestrowaniem stanów środowiska (por. np.: Thompson 2007; Stewart, Gapenne, Di Paolo 2010). Kwestia relacji, jakie zachodzą między koncepcją poznania ucieleśnionego, koncepcją poznania usytuowanego oraz enaktywizmem jest osobnym zagadnieniem teoretycznym, którego nie sposób tu omówić. Co jednak godne zaznaczenia, to fakt, że odrzucając różne założenia „tradycyjnej" kognitywistyki, zwolennicy tych podejść często odrzucają także reprezentacje jako eksplanacyjnie wartościowe narzędzia. Istnieją co prawda próby pogodzenia idei poznania jako ucieleśnionego i nastawionego na działanie z ideą, że poznanie opiera się na reprezentacjach (Grush 1997, 2004; Bickhard 2004a, 2004b, 2009; Anderson, Rosenberg 2008; Goldman 2012). Niektóre z tych prób zostaną omówione na kartach tej książki. Jednak nawet mimo tych starań, ogólny intelektualny „klimat" panujący wokół nowszych podejść do naukowego badania poznania jest często wyraźnie antyreprezentacjonistyczny.

Z punktu widzenia prowadzonych tu rozważań antyreprezentacjonizm jest interesujący nie tyle jako kategoria *opisująca* panujące w kognitywistyce trendy teoretyczne i badawcze (antyreprezentacjonizm deskryptywny), co raczej jako stanowisko *preskryptywne*. Tak sformułowany antyreprezentacjonizm opiera się na tezie, że reprezentacje mentalne powinny zniknąć z eksplanacyjnego repozytorium kognitywistyki. Poprawne wyjaśnienia w kognitywistyce nie powinny postulować wewnętrznych reprezentacji środowiska. Sama idea, że relację między systemem poznawczym a jego środowiskiem należy rozpatrywać w kategoriach reprezentacyjnych, jest zasadniczą pomyłką. Opisany tu artykuł van Geldera (1995) to bardzo klarowny i otwarty wyraz antyreprezentcjonizmu w takim preskryptywnym znaczeniu, ponieważ autor ten pokazuje, że (oraz dlaczego) badacze powinni odrzucić „standardową", reprezentacjonistyczną praktykę eksplanacyjną. Choć van Gelder postuluje też określoną strategię alternatywną – sugeruje pewną pozytywną koncepcję dotyczącą tego, w którym kierunku powinna zdążać kognitywistyka po odrzuceniu reprezentacjonistycznego „bagażu" teoretycznego – to w przyjętym tu rozumieniu antyreprezentacjonizm pojmuje

się jako czysto *negatywne* stanowisko (nieniosące ze sobą z konieczności żadnych konkretnych wskazówek, czym powinniśmy zastąpić reprezentacjonizm). Być może każdy antyreprezentacjonista jest intelektualnie zobligowany do zaproponowania takiej alternatywy teoretycznej, jednak żadna z możliwych propozycji nie jest związana z antyreprezentacjonizmem definicyjnie.

W ramach tej pracy będę też przyjmował – analogicznie do tego, jak to było w przypadku reprezentacjonizmu – że antyreprezentacjonizm rozumiany jako zalecenie dotyczące (pożądanej, poprawnej, uzasadnionej) praktyki *eksplanacyjnej* kognitywistów jest ściśle powiązany z tezą dotyczącą tego, jak *rzeczywiście* działa system poznawczy. Będę zatem przyjmował, że antyreprezentacjonizm głosi, iż powinniśmy unikać reprezentacji w naszych wyjaśnieniach, ponieważ te rzeczywiście nie pełnią żadnej roli funkcjonalnej w aktywności systemu poznawczego. Raz jeszcze należy więc zaznaczyć, że odrzucam tezę Chemero (2009: 67–83), aby traktować (anty)reprezentacjonizm w rozumieniu epistemicznym oraz metafizycznym jako oddzielne tezy, z których każda może być broniona niezależnie od drugiej. W przyjętej tu przeze mnie perspektywie epistemiczny sukces wyjaśniania działania systemu poznawczego bez odwoływania się do reprezentacji wynika z faktów dotyczących niezależnej od obserwatora, wewnętrznej organizacji tego systemu.

1.1.3. Spór reprezentacjonizm–antyreprezentacjonizm: poziom przedmiotowy i metaprzedmiotowy

Kolejną analitycznie ważną na potrzeby tej pracy dystynkcją jest rozróżnienie na *przedmiotowy* oraz *metaprzedmiotowy* wymiar sporu o status eksplanacyjny reprezentacji mentalnych w kognitywistyce. Na poziomie przedmiotowym spór między reprezentacjonistami a antyreprezentacjonistami dotyczy tego, czy system poznawczy jest systemem reprezentacyjnym, to znaczy takim, którego działanie należy wyjaśniać za pomocą reprezentacji[9]. Wedle jednej strony repre-

[9] W dalszej części tej pracy za pomocą terminów „reprezentacjonizm" i „antyreprezentacjonizm" będę oznaczać jedynie preskryptywne wersje stanowisk opisanych w poprzednich sekcjach.

zentacje powinny stanowić jedno z zasadniczych narzędzi służących do wyjaśniania interesujących kognitywistykę zjawisk. Bez postulowania wewnętrznych reprezentacji środowiska nie wyjaśnimy większości lub wręcz żadnego spośród fenomenów stanowiących eksplanand tej dyscypliny naukowej. Druga strona sporu twierdzi z kolei, że poprawne wyjaśnienia w kognitywistyce mogą (oraz powinny) obyć się bez postulowania reprezentacji. Jakkolwiek należy rozumieć relację między systemem poznawczym a środowiskiem, pewne jest, że powinniśmy zrezygnować z pojmowania tej relacji jako zapośredniczonej reprezentacjami.

Powyższy opis tego, o co toczy się spór reprezentacjonizm–antyreprezentacjonizm na poziomie przedmiotowym, jest rzecz jasna dość ogólny i szkicowy. Pomija on niektóre potencjalnie ważne różnice między rozmaitymi sformułowaniami obu tych stanowisk. Przede wszystkim należałoby uzupełnić powyższą charakterystykę, wprowadzając odróżnienie reprezentacjonizmu i antyreprezentacjonizmu w wersji globalnej od bardziej lokalnych wersji tych stanowisk. Przy ujęciu globalnym przedmiotem sporu byłaby kwestia możliwości całkowitego odrzucenia pojęcia reprezentacji jako narzędzia eksplanacyjnego kognitywistyki. Antyreprezentacjoniści byliby w takim wypadku zobligowani do przyjęcia bardzo mocnej tezy, zgodnie z którą pojęcie reprezentacji nie jest przydatne do zrozumienia któregokolwiek fenomenu stanowiącego eksplanandum dla kognitywistyki. Oznaczałoby to fundamentalną rekonceptualizację tradycyjnych dla tej dyscypliny założeń dotyczących natury systemu poznawczego oraz procesów poznawczych. Analogicznie, globalna wersja reprezentacjonizmu opierałaby się na mocnej tezie, że wszelkie zjawiska poznawcze powinny być wyjaśniane przez postulowanie reprezentacji mentalnych.

W alternatywnym, „lokalnym" rozumieniu przedmiot sporu między zwolennikami reprezentacjonizmu i ich oponentami zawężałby się do tego czy innego, konkretnego eksplanandum. Innymi słowy, w tej sytuacji mielibyśmy do czynienia z całym szeregiem sporów dotyczących konkretnych zjawisk poznawczych wymagających wyjaśnienia. Na przykład jako kwestię podlegającą dyskusji można by potraktować to, czy konkretne zjawisko związane z poznaniem

społecznym – chociażby zdolność do podzielania uwagi z innymi – wymaga wyjaśnienia w kategoriach reprezentacyjnych. Jednak decyzja w sprawie preferowanego wyjaśnienia tego akurat eksplanandum niekoniecznie niosłaby ze sobą jakiekolwiek konsekwencje, jeśli chodzi o preferowany sposób wyjaśniania jakiegoś innego zjawiska, na przykład zdolności do nabycia języka. Reprezentacjonizm mógłby w takim razie „wygrywać" na pewnych frontach, a stanowisko przeciwne – na innych. Nie mielibyśmy jednak do czynienia z sytuacją, w której „zwycięzca bierze wszystko".

Warto również podkreślić, że w ramach przyjętych tu założeń – precyzyjniej wyrażę i uzasadnię je w dalszej części tej pracy (por. zwłaszcza, sekcja 3.1.2) – spór o preferowany sposób *wyjaśniania* systemu poznawczego nie jest niezależny od sporu o *naturę* tego systemu. Zgodnie z przyjętą tu przeze mnie perspektywą sukces eksplanacyjny w kognitywistyce okazuje się częściowo uwarunkowany rzeczywistą, niezależną od obserwatora przyczynową/funkcjonalną architekturą systemu poznawczego. Spór o eksplanacyjny status reprezentacji jest jednocześnie sporem o to, czy i w jakim zakresie system poznawczy to system reprezentacyjny, to znaczy – system wykorzystujący reprezentacje w swoim działaniu.

Przejdźmy teraz do problemu statusu eksplanacyjnego reprezentacji w naukach kognitywnych w jego *metaprzedmiotowym* wymiarze. Wszystkie dotychczasowe uwagi czynią założenie, że istnieje powszechnie znana i akceptowana wiedza o tym, czym są wyjaśnienia reprezentacyjne i czym się różnią od wyjaśnień niereprezentacyjnych. Jak się wydaje, spór między zwolennikami i przeciwnikami reprezentacjonizmu ma sens tylko wtedy, gdy istnieje konsensus dotyczący właśnie tego metaprzedmiotowego zagadnienia, które nie dotyczy już pożądanej strategii wyjaśniania, ale tego, *na czym ta strategia właściwie polega* i czym dokładnie różni się od strategii alternatywnej.

Istotę oraz wagę metaprzedmiotowego wymiaru opozycji reprezentacjonizm–antyreprezentacjonizm można zilustrować, wracając do dyskusji nad zaproponowanym przez van Geldera wyjaśnieniem silnika Watta. Silnik ten jest stosunkowo prostym układem mechanicznym. Jak się wydaje, nieformalna charakterystyka tego urządze-

nia, dodatkowo uzupełniona o charakterystykę w kategoriach dynamicznych, mówi nam o nim wszystko, co powinniśmy wiedzieć, aby zrozumieć jego działanie. Okazuje się jednak, że w literaturze wcale nie istnieje konsensus co do tego, czy ten prosty układ mechaniczny rzeczywiście nie wykorzystuje reprezentacji. William Bechtel (1998) zaproponował reprezentacyjne wyjaśnienie działania silnika Watta. Zgodnie z propozycją Bechtela rola reprezentacji w dowolnym systemie fizycznym polega na „zastępowaniu" jakiegoś obiektu, stanu rzeczy czy zdarzenia. Reprezentacje pełnią tego rodzaju rolę dzięki temu, że niosą informację o tym, co reprezentowane. Jednak kodowanie informacji nie wystarcza dla bycia reprezentacją. Rola tej ostatniej polega raczej na udostępnianiu niesionej przez nią informacji tym komponentom systemu, których poprawne funkcjonowanie wymaga odpowiedniej koordynacji z reprezentowanym obiektem, stanem rzeczy czy zdarzeniem. Zgodnie z ujęciem Bechtela mamy do czynienia z systemem reprezentacyjnym, gdy w grę wchodzą trzy elementy: (1) zewnętrzne zdarzenie, stan rzeczy czy obiekt; (2) komponent systemu kodujący informację o (1); (3) komponent, którego poprawne funkcjonowanie wymaga odpowiedniej koordynacji z (1), przy czym ta koordynacja jest możliwa dzięki temu, że ten komponent używa informacji, która jest kodowana przez (2). Pełnoprawnie reprezentacyjne wyjaśnienie pokazuje zatem, że badany system realizuje określone funkcje dzięki wykorzystywaniu wewnętrznych, informacyjnych „pośredników" między wewnętrznymi komponentami a zewnętrznymi stanami rzeczy (zdarzeniami, obiektami).

Zdaniem Bechtela (1998) przyjęcie proponowanej przez niego perspektywy pozwala w jasny i przekonujący sposób pokazać, że (wbrew van Gelderowi) silnik Watta powinien zostać uznany za system korzystający z reprezentacji; tym samym powinniśmy wyjaśniać jego działanie, odwołując się do wewnętrznych reprezentacji. Bechtel zwraca uwagę na fakt, że kąt rozstawu ramion regulatora Watta systematycznie współzmienia się z prędkością koła zamachowego, dzięki czemu ten pierwszy koduje informację o tej drugiej. Ta informacja jest wykorzystywana przez zawór w cylindrze, aby zmienić panujące tam ciśnienie na takie, które zapewnia odpo-

wiednią prędkość ruchu koła zamachowego. Tym samym regulator Watta reprezentacyjnie „zastępuje" prędkość koła zamachowego dla zaworu doprowadzającego parę do cylindra. Prędkość koła zamachowego jest reprezentowana przez kąt rozstawu ramion regulatora, a zawór w cylindrze korzysta z tej reprezentacji, aby odpowiednio dostosować swoje funkcjonowanie. Mamy zatem do czynienia z reprezentowanym stanem rzeczy, komponentem stanowiącym nośnik informacji o tym stanie rzeczy oraz komponentem korzystającym z tej informacji do odpowiedniej koordynacji własnego działania względem tego stanu rzeczy. Silnik Watta spełnia postawione przez Bechtela warunki bycia systemem reprezentacyjnym.

Bechtel kontruje także dodatkowe, opisane w poprzednim podrozdziale, argumenty przywoływane przez van Geldera (1995) na rzecz antyreprezentacjonistycznego wyjaśnienia silnika Watta. Po pierwsze, przyznaje on van Gelderowi, że sama współzmienność nie wystarczy do zapewnienia czemuś statusu reprezentacji. Jednakże zgodnie z jego propozycją współzmienność między regulatorem Watta a prędkością koła zamachowego jest *wykorzystywana* przez zawór w cylindrze. Dopiero ten fakt – że regulator jest wykorzystywany jako reprezentacja – stanowi warunek wystarczający dla tego, aby stosowanie wyjaśnienia reprezentacyjnego było uzasadnione. Po drugie, inspirując się koncepcjami Ruth Millikan (1984), Bechtel argumentuje, że chociaż współzmienność między regulatorem Watta a prędkością koła zamachowego nie występuje zawsze[10], to zachodzenie tej kowariancji wcale nie jest wymagane do tego, aby można było utrzymywać, że jedno reprezentuje drugie. Mówiąc ogólnie, funkcja elementu może polegać na reprezentowaniu czegoś czegoś nawet wtedy, kiedy *de facto* odpowiednia współzmienność zachodzi rzadko, a nawet nie zachodzi w ogóle. Analogicznie, regulator Watta może funkcjonować jako reprezentacja, nawet jeśli kąt rozstawu ramion jest często opóźniony w stosunku do prędkości koła zamachowego. Po trzecie wreszcie, Bechtel przyznaje, że sprzężenie między kątem rozstawu ramion regulatora a prędkością koła zamachowe-

[10] Kąt rozstawu ramion regulatora Watta oraz prędkość koła zamachowego są bowiem często rozregulowane.

go stanowi relację zbyt subtelną, żeby wyrazić ją za pomocą niektórych pojęć reprezentacji. Postuluje on jednak, że możemy zachować ogólną ideę reprezentowania jako „zastępowania", a jednocześnie uzupełnić ją twierdzeniem, iż reprezentowanie może być oparte na dynamicznym sprzężeniu. Tym samym powinniśmy wedle Bechtela nie tyle odrzucić pojęcie reprezentacji, co raczej je zmodyfikować w taki sposób, aby nie zakładało ono z góry, że reprezentowanie to statyczna relacja, zasadniczo niekompatybilna z podejściem dynamicznym.

Nie chodzi tu w tej chwili o to, aby opowiedzieć się za jedną ze stron sporu dotyczącego poprawnego wyjaśnienia silnika Watta. Rzecz też nie w tym, aby sugerować, że dylemat ten jest nierozstrzygalny. W dalszej części tej książki (w rozdziale 4) będę bronić koncepcji wyjaśniania reprezentacyjnego, która jest wyraźnie spokrewniona z tą zaproponowaną przez Bechtela, jednak zarazem różna od stanowiska tego autora na tyle, by przynieść zupełnie inne rozstrzygnięcie dotyczące statusu silnika Watta. Na obecnym etapie rozważań chcę jednak zwrócić uwagę na inny aspekt sporu Bechtela z van Gelderem. Otóż wychodzi w nim na jaw rzecz, która, jak się wydaje, jest symptomatyczna dla prowadzonej współcześnie dyskusji nad reprezentacjami i ich eksplanacyjnym statusem. Chodzi o przykład sytuacji, gdy dwóch autorów fundamentalnie nie zgadza się co do tego, czy bardzo prosty układ mechaniczny wymaga wyjaśnienia za pomocą reprezentacji. Mechanizm, na podstawie którego działa silnik Watta, jest nieskomplikowany; jego nieformalny opis daje nam poczucie niemal kompletnego zrozumienia, jak działa ten system. Co więcej, obie strony sporu do pewnego stopnia zgadzają się co do tego, jak on działa. Mimo to nawet w tym pozornie łatwym przypadku dwóch kompetentnych autorów może spierać się o to, czy uzasadnione jest wyjaśnianie badanego systemu w kategoriach reprezentacyjnych. Pokazuje to, że przedmiotowy spór o reprezentacjonizm prowadzi się w sytuacji, w której między obydwiema stronami nie ma konsensusu dotyczącego metaprzedmiotowego problemu dotyczącego tego, na czym dokładnie polega wyjaśnienie pewnego fenomenu za pomocą reprezentacji. Nie istnieje konsensus co do tego, czym jest wyjaśnianie za pomocą reprezentacji. Pim Haselager,

Andre de Groot i Hans van Rappard (2003) wyrażają tę myśl, stwierdzając, że pojęciu reprezentacji stosowanemu w naukach kognitywnych brakuje satysfakcjonującej operacjonalizacji. Dyskusje o reprezentacyjnej lub niereprezentacyjnej naturze określonych systemów w większości są oparte na niejasnych, nie zawsze podzielanych powszechnie intuicjach i presupozycjach. Dlatego kognitywiści nie mogą obecnie udzielić jasnej, powszechnie akceptowalnej odpowiedzi na pytanie o to, jakie obserwowalne własności systemu poznawczego gwarantowałyby, że wyjaśnienie działania tego systemu w kategoriach reprezentacyjnych byłoby poprawne.

Zauważmy, że opisany tu spór między van Gelderem i Bechtelem jest otwartą dyskusją na temat tego, czym są wyjaśnienia reprezentacyjne. Nawet jeśli dyskusja ta pokazuje, jak bardzo dwóch autorów może nie zgadzać się co do sposobu wyjaśniania prostego układu mechanicznego, to obydwie strony są przynajmniej zaangażowane w próby uzasadnienia *explicite* zajmowanych przez siebie pozycji. Sytuacja ta nie jest charakterystyczna dla ogółu kognitywistyki. Praktyce posługiwania się terminem „reprezentacja mentalna" nie towarzyszą na ogół próby wyrażenia wprost, co termin ten w określonym kontekście ma oznaczać. Ramsey (2007: 7) przytacza słowa Stephena Palmera, wedle którego „jako psychologowie poznawczy nie rozumiemy tak naprawdę pojęcia reprezentacji. Postulujemy reprezentacje, mówimy o nich, spieramy się o nie, próbujemy znaleźć dowody na ich istnienie, lecz zasadniczo ich nie rozumiemy" (Palmer 1979: 259[11]). W przypadku Ramseya (2007) diagnoza dotycząca takiego pojęciowego zamieszania stanowi punkt wyjścia do obrony stanowiska, które można nazwać „antyreprezentacjonizmem rewizyjnym". Ramsey nie usiłuje pokazywać, że nowe, „lepsze" teorie procesów poznawczych są antyreprezentacjonistyczne, lecz próbuje uzasadnić tezę, że kognitywiści już od jakiegoś czasu posługują się pojęciem reprezentacji tak liberalnie i niejasno, że przy bliższym spojrzeniu okazuje się, iż nie odwołują się oni do reprezentacji w żadnym eksplanacyjnie wartościowym sensie. Mówiąc ina-

[11] Wszystkie cytaty pochodzące z artykułów i książek nieprzełożonych na język polski zostały przetłumaczone przez autora.

czej, spora część wyjaśnień proponowanych przez kognitywistów tylko pozornie odwołuje się do reprezentacji. Kognitywistyka znajduje się według Ramseya (2007: 188–235) w „postreprezentacjonistycznym" stadium swojej historii już od czasu powstania i upowszechnienia koneksjonizmu[12].

Powyższe uwagi pozwalają zrozumieć, że metaprzedmiotowy wymiar sporu o reprezentacje ma dla kognitywistyki znaczenie fundamentalne. Bez podjęcia próby jego rozwiązania trudno liczyć na konkluzywne rozwiązanie *przedmiotowego* sporu między reprezentacjonistami a antyreprezentacjonistami. Jeśli nie chcemy poruszać się „po omacku" na przedmiotowym poziomie, potrzebujemy dobrej koncepcji tego, czym jest wyjaśnianie reprezentacyjne.

Podsumowując, metaprzedmiotowy problem eksplanacyjnego statusu reprezentacji mentalnych w kognitywistyce wyznaczają dwa pytania:

> **1.** Co czyni wyjaśnienia badanego systemu wyjaśnieniami reprezentacyjnymi? Co odróżnia je od wyjaśnień niereprezentacyjnych?
> **2.** Jakie własności badanego systemu sprawiają, że system ten jest systemem reprezentacyjnym?

Pytanie 1 kładzie nacisk na epistemiczny aspekt problemu, a pytanie 2 – na jego metafizyczny aspekt. Tak jak w przypadku problemu przedmiotowego, tak i tu oba pytania należy rozpatrywać jako alternatywne sformułowania jednego problemu. Jeśli poprawnie wyjaśniamy działanie danego systemu w kategoriach reprezentacyjnych, to znaczy, że ten system rzeczywiście posługuje się reprezentacjami. Zarazem jeśli system ten posługuje się reprezentacjami, to reprezentacyjne wyjaśnienie jego działania będzie poprawne.

[12] Z tym twierdzeniem będę polemizować w rozdziale 4 (podrozdział 4.3).

1.2. Spór o reprezentacjonizm a naturalizowanie intencjonalności

Wydaje się, że kiedy filozofowie umysłu używają terminu „problem reprezentacji mentalnych", na ogół nie mają na myśli problemu statusu eksplanacyjnego reprezentacji w kognitywistyce. „Problem reprezentacji" w rozumieniu najbardziej rozpowszechnionym we współczesnej filozofii umysłu jest zagadnieniem przynajmniej *prima facie* osobnym i autonomicznym względem rozstrzygnięć dotyczących pożądanej czy uzasadnionej praktyki eksplanacyjnej w ramach kognitywistyki. Dla większości filozofów najbardziej doniosły problem związany z pojęciem reprezentacji mentalnych nie jest bowiem *metodologicznym* zagadnieniem powstającym w obrębie kognitywistyki, lecz *metafizycznym* pytaniem o to, jak jest możliwe, żeby intencjonalność cechująca stany mentalne mogła być egzemplifikowana w świecie opisywanym i wyjaśnianym – w sposób pozbawiony kategorii intencjonalnych – przez fizykę i inne nauki szczegółowe. Tym samym kanoniczny „problem reprezentacji" dla filozofii umysłu stanowi kwestia, którą nazywa się na ogół problemem *naturalizacji intencjonalności*.

Źródłem problemu naturalizacji intencjonalności jest rozpoznanie, że stany mentalne, które przypisujemy sobie oraz innym w toku codziennych interakcji społecznych – stany, które wyznaczają nasze podstawowe, przednaukowe rozumienie umysłu – są intencjonalne, to znaczy mają pewną treść, są zawsze „o czymś". Stanami tymi są przekonania, pragnienia, intencje, nadzieje czy oczekiwania, słowem – postawy propozycjonalne. Na ogół przyjmuje się też, że przypisywanie tym stanom intencjonalności jest równoznaczne z twierdzeniem, że są one *reprezentacjami*. Przekonanie, że pada deszcz – przekonanie o treści: „Pada deszcz" – reprezentuje stan rzeczy polegający na tym, że pada deszcz; posiada ono warunki poprawności lub prawdziwości, które są spełnione, o ile rzeczywiście pada deszcz. Jak wspomniałem we wstępie, aparat pojęciowy odwołujący się do przekonań, pragnień i innych postaw propozycjonalnych – służący ludziom do przewidywania i wyjaśniania własnych i cudzych działań w toku codziennych interakcji – nazywa się na ogół „psychologią potoczną". Stany inten-

cjonalne postulowane w ramach psychologii potocznej nie tylko mają treść, ale też pełnią określone role funkcjonalne/przyczynowe. Każda postawa propozycjonalna ma pewien „profil" funkcjonalny (przyczynowy), określający jej przyczynowe interakcje z działaniami oraz innymi postawami propozycjonalnymi. Te funkcjonalne role pełnione przez postawy propozycjonalne względem działań i innych stanów mentalnych są z kolei determinowane właśnie przez ich treść. Przekonania i pragnienia wpływają na inne postawy oraz kształtują działania w pewien określony sposób dlatego, że mają taką, a nie inną treść. Twierdząc, że Jan kupił samochód, ponieważ pragnął zakupić samochód, mamy między innymi na myśli, że Jan postąpił tak, a nie inaczej ze względu na treść jego pragnienia; pragnienie pokierowało jego działaniem w określony sposób dlatego, że miało taką, a nie inną treść.

Problem naturalizacji intencjonalności sprowadza się do pytania: Jak stany intencjonalne mogą stanowić część świata fizycznego? Jak stany czysto biologicznie czy fizykalnie opisanego systemu poznawczego mogą egzemplifikować własność „bycia o", reprezentowania czy posiadania treści intencjonalnej? Dla większości filozofów umysłu udzielenie odpowiedzi na te pytania jest równoznaczne z koniecznością wyrażenia, w kategoriach czysto naturalistycznych i nieintencjonalnych (fizykalnych, biologicznych, funkcjonalnych czy informacyjnych), warunków wystarczających dla tego, byśmy mogli zidentyfikować pewien stan wewnętrzny organizmu (systemu poznawczego, mózgu) z postawą propozycjonalną o określonej treści intencjonalnej. Jeśli intencjonalność stanowi część świata naturalnego, to istnieją wewnętrzne stany podmiotów, które rzeczywiście przyczynowo kształtują działania, a które egzemplifikują czysto naturalne własności sprawiające, że stany te są stanami intencjonalnymi. Kiedy poprawnie przypisujemy komuś przekonanie o określonej treści, w rzeczywistości odnosimy się do realnego, naturalnego i przyczynowo efektywnego stanu tej osoby, który posiada tę treść i który przyczynowo odpowiada za działania (w tym działania inferencyjne) tej osoby. Projekt naturalizacji polega na wskazaniu owych naturalnych i nieintencjonalnych własności, czyniących ten stan intencjonalnym.

Nie będę tu szczegółowo rekonstruować podejmowanych na przełomie ostatnich dziesięcioleci prób naturalizacji intencjonalności. Warto jednak przedstawić choćby ich zarys. Poszczególni filozofowie próbowali naturalizować intencjonalność (dokładniej: treść intencjonalną), odwołując się do pojęcia informacji (Dretske 1981; 1986; 1988), funkcji biologicznych (Millikan 1984; Papineau 1987), asymetrycznych zależności przyczynowych (Fodor 1987), czy też redukując treść stanów intencjonalnych do ogółu pełnionych przez nie ról funkcjonalnych/przyczynowych (Block 1986; Harman 1987). Teorie te stawały przez licznymi szczegółowymi wyzwaniami, takimi jak chociażby: (1) wyjaśnienie w naturalistyczny sposób, jak możliwe są reprezentacje fałszywe; (2) wyjaśnienie cechującej postawy propozycjonalne intensjonalności[13]; (3) pokazanie, w jaki sposób jest zapewniona odpowiednia „ziarnistość" treści[14]; (4) udzielenie odpowiedzi na pytanie o to, czy treść jest determinowana przez własności wewnętrzne, czy też przez relacje zachodzące między wewnętrznymi stanami systemu poznawczego a środowiskiem zewnętrznym (spór internalizm–eksternalizm w sprawie treści mentalnej).

Pamiętajmy jednak, że posiadanie treści intencjonalnej nie wystarcza do bycia stanem intencjonalnym (rozumianym jako postawa propozycjonalna). Postawy propozycjonalne mają także funkcjonalne własności: w określony sposób wpływają one na działania, w tym działania inferencyjne (rozumowania). Co więcej, role funkcjonalne postaw propozycjonalnych są wyznaczane przez treść tych postaw. Pochopne byłoby wykorzystywanie zamiennie terminów „naturalizacja treści" i „naturalizacja intencjonalności". Naturalizacja treści jest warunkiem koniecznym, lecz niewystarczającym udanej naturalizacji intencjonalności. Każda naturalistyczna teoria intencjonalności aspirująca do miana kompletnej musi pokazywać, że

[13] To jest wyjaśnienie, jak mogą istnieć koekstensjonalne, lecz nieidentyczne (bo treściowo różne) postawy propozycjonalne (bardziej obrazowo: jak w naturalistyczny sposób wyjaśnić, czym różnią się koekstensjonalne przekonania o Marilyn Monroe i Normie Jean Mortenson).

[14] To jest pokazanie w *stricte* naturalistyczny sposób, że treści intencjonalne mogą być odpowiednio „szczegółowe" (na przykład, że jest możliwe odróżnienie przekonań o królikach od przekonań o nierozłączonych częściach królików).

postulowane przez nią stany („znaturalizowane" postawy propozycjonalne) odpowiadają pod względem *funkcjonalnym* stanom intencjonalnym psychologii potocznej. „Znaturalizowane" przekonanie, że pada deszcz, powinno zatem nie tylko (1) posiadać treść, że pada deszcz, ale też (2) sprawiać (w określonych okolicznościach), że jego posiadacz zrezygnuje z wyjścia z domu, albo że wyjdzie z domu, lecz zabierze ze sobą parasol, że odpowie twierdząco na pytanie o to, czy pada deszcz i tak dalej. Wreszcie (3) przekonanie to powinno pełnić role opisane w (2) właśnie dlatego, że posiada treść wymienioną w (1). Teorie wykorzystujące elementy teleologiczne (por. np.: Millikan 1984; Dretske 1986, 1988) oraz teorie wprost redukujące intencjonalność do ról funkcjonalnych (por. np.: Block 1986; Harman 1987) wydawały się szczególnie dobrze dostosowane do uczynienia zadość tym wyzwaniom teoretycznym.

Jak już zaznaczyłem, może wydawać się *prima facie*, że problem naturalizacji intencjonalności i problem eksplanacyjnego statusu reprezentacji w kognitywistyce to osobne, autonomiczne zagadnienia. Zachodzące między nimi różnice są wymienione w tabeli 1. Po pierwsze, problem statusu eksplanacyjnego reprezentacji należy do metodologicznych zagadnień kognitywistyki, natomiast zagadnienie naturalizacji intencjonalności przynależy do (naturalistycznej) metafizyki umysłu. Potrzeby teoretyczne stojące u podstaw obydwu problemów są zatem różne. Po drugie, problem statusu eksplanacyjnego reprezentacji skupia się na stanach i strukturach, które na ogół zasadniczo różnią się od postaw propozycjonalnych pod względem przypisywanych im własności semantycznych i funkcjonalnych. Wykorzystując rozróżnienie, do którego jeszcze powrócę w tej książce (w rozdziale 5), można powiedzieć, że spór reprezentacjonizm–antyreprezentacjonim w kognitywistyce dotyczy stanów czy struktur o charakterze *subosobowym*, natomiast przedmiotem problemu naturalizacji intencjonalności są *osobowe* stany przypisywane podmiotom intencjonalnym. Po trzecie, przedstawiciele obu stron sporu między reprezentacjonistami a antyreprezentacjonistami w kognitywistyce są przede wszystkim zainteresowani reprezentacjami jako eksplanansami, czyli *narzędziami*, za pomocą których wyjaśnia

się zjawiska[15]. W projekcie naturalizacji intencjonalności reprezentacje (postawy propozycjonalne) są traktowane z kolei jako *przedmioty* wyjaśniania, czyli jako eksplananda[16].

Czy jednak pełna autonomia obydwu omawianych „problemów reprezentacji" jest do utrzymania? Czy potencjalne rozwiązania jednego z nich nie niosą ze sobą żadnych implikacji dla prób rozwiązania drugiego, i *vice versa*? Przy bliższym spojrzeniu okazuje się, że utrzymywanie takiego „separacjonistycznego" stanowiska wydaje się kontrowersyjne. Załóżmy, że któremuś z przedstawicieli programu naturalizacji udaje się odnieść sukces. Przedstawia on dobrze

[15] Teza ta jest pewnym uproszczeniem. Kognitywiści często twierdzą, że dane zjawisko kognitywne należy wyjaśnić reprezentacyjnie, jednak nie wiedzą dokładnie, jaką postać przyjmują postulowane przez nich reprezentacje (czym one są, gdzie w mózgu są zlokalizowane i tak dalej). Czynią oni więc przedmiotem swoich dociekań właśnie naturę tych ostatnich (por. Ramsey 2007: 34–36). Chcą oni *wyjaśnić* reprezentacje, a nie tylko wyjaśnić coś *za pomocą* reprezentacji. Tym samym nie jest tak, że reprezentacje nigdy nie są traktowane w kognitywistyce jako eksplananda. Jak jednak zauważa Ramsey (2007), nawet w takich wypadkach okazuje się, że (1) reprezentacje jako eksplananda odgrywają podwójną rolę, ponieważ mają też ostatecznie pełnić funkcję eksplanansu, to jest służyć do wyjaśnienia jakiegoś innego zjawiska; (2) do kryteriów poprawności wyjaśnienia reprezentacji (rozumianej jako eksplanandum) należy to, czy okaże się ona potencjalnie użyteczna jako *eksplanans*. Traktowanie reprezentacji jako eksplanandów w kognitywistyce jest w dużym stopniu pochodne względem traktowania ich jako eksplanansów. Charakteryzowanie reprezentacjonizmu i antyreprezentacjonizmu jako stanowisk dotyczących przede wszystkim eksplanansów wydaje się zatem uzasadnione.

[16] Należy poczynić tu dwie uwagi. Po pierwsze, termin „wyjaśnienie" jest w tym kontekście używany w bardzo ogólny i teoretycznie niezobowiązujący sposób. „Wyjaśnieniem" w tym znaczeniu jest dowolny twór epistemiczny (teoria, hipoteza, model i tak dalej), który umożliwia zrozumienie danego fenomenu. W tym ogólnym sensie przedstawiciele projektu naturalizacji są zainteresowani wyjaśnieniem, jak możliwe jest – mówiąc obrazowo – istnienie przekonań i innych postaw propozycjonalnych w świecie fizycznym. Po drugie, należy pamiętać, że teza, iż reprezentacje w programie naturalizacji intencjonalności pełnią rolę eksplanandów, to pewna idealizacja. Jak się bowiem okazuje, to, jaki użytek można uczynić (lub nie) z postaw propozycjonalnych jako eksplanansów wcale nie jest bez wpływu na rozwiązania problemu naturalizacji intencjonalności. Na przykład niektórzy eliminatywiści – ujmując rzecz w sporym skrócie – negują istnienie stanów intencjonalnych na podstawie tego, że nie są one dobrym narzędziem eksplanacyjnym (eksplanansem) dla kognitywistów.

Tabela 1. Zestawienie różnic zachodzących między problemem naturalizacji intencjonalności a problemem statusu eksplanacyjnego reprezentacji w kognitywistyce

Rodzaj problemu	Problem naturalizacji intencjonalności	Problem statusu eksplanacyjnego reprezentacji w kognitywistyce
Czego dotyczy problem?	Jakie są czysto naturalne własności, które może egzemplifikować wewnętrzny stan systemu poznawczego, a których posiadanie wystarcza do tego, by stan ten był identyczny z postawą propozycjonalną o określonej treści?	Czy kognitywiści potrzebują pojęcia reprezentacji do realizacji stawianych sobie celów eksplanacyjnych? Czy dobre (poprawne, najlepsze spośród dostępnych) wyjaśnienia odwołują się do wewnętrznych reprezentacji? Czy (lub w jakim zakresie) system poznawczy jest systemem reprezentacyjnym?
Jakiego rodzaju reprezentacje są przedmiotem problemu?	Osobowe, intencjonalne stany podmiotów. Postawy propozycjonalne, takie jak przekonania, pragnienia, intencje, obawy, nadzieje, oczekiwania i tak dalej.	Subosobowe stany i struktury (biologiczne, funkcjonalne/obliczeniowe), które mogą pełnić funkcję reprezentacji: struktury danych, komórki receptorowe, wzorce połączeń synaptycznych w sieci koneksyjnej, emulatory układu mięśniowo-szkieletowego i tak dalej.
W jakiej roli występują reprezentacje?	Reprezentacje mentalne jako (przede wszystkim) eksplanandum.	Reprezentacje mentalne jako (przede wszystkim) eksplanans.

uargumentowane, spełniające wszelkie możliwe oczekiwania teoretyczne stanowisko w sprawie czysto naturalnych warunków wystarczających do tego, aby pewien stan biologiczny lub funkcjonalny/obliczeniowy mógł być utożsamiony z postawą propozycjonalną. Wyobraźmy sobie jednak, że najlepsze dostępne teorie czy modele z zakresu nauk kognitywnych pokazują, iż w systemie poznawczym *nie istnieje nic, co spełnia te warunki*. Inaczej mówiąc, wiedza z zakresu kognitywistyki okazuje się zasadniczo niekompatybilna z ideą, że nasze „znaturalizowane" postawy propozycjonalne odgry-

wają rolę w aktywności systemu poznawczego. Znaczy to, że zgodnie z najlepszą dostępną naukową taksonomią stanów mentalnych: (1) stany wyróżniane w ramach tej taksonomii nie odpowiadają postawom propozycjonalnym pod względem własności semantycznych (mają inne treści), lub (2) stany te nie odpowiadają postawom propozycjonalnym pod względem pełnionych przez nie ról funkcjonalnych, lub (3) stany te odpowiadają semantycznie i funkcjonalnie postawom propozycjonalnym, lecz (w przeciwieństwie do postaw propozycjonalnych) ich własności semantyczne nie determinują ich ról funkcjonalnych, lub (4) stany te w ogóle nie są reprezentacjami. Przypadki (1), (2) i (3) to sytuacje, w których kognitywistyka co prawda postuluje reprezentacje, lecz są one zasadniczo różne od tych postulowanych w ramach psychologii potocznej. Przypadek (4) to sytuacja, w której kognitywistyka obywa się całkowicie bez postulowania wewnętrznych reprezentacji. Co jeśli któraś z tych możliwości zachodzi? Jakie konsekwencje miałby taki stan rzeczy dla projektu naturalizacji intencjonalności?

Wielu filozofów bardziej lub mniej otwarcie przyjmuje, że zajście jednej z wymienionych wyżej możliwości miałoby zasadnicze konsekwencje dla filozoficznych prób „umieszczenia" intencjonalności w porządku naturalnym. Wyrazem takiego założenia jest choćby często przytaczane przekonanie Jerry'ego Fodora, że kognitywistyka „zrehabilituje" psychologię potoczną jako (w przybliżeniu) prawdziwą koncepcję tego, jak działa umysł, a jeśli tak się nie stanie, będziemy świadkami największej katastrofy intelektualnej w dziejach gatunku ludzkiego – okaże się bowiem, że nasza psychologia przekonań i pragnień jest fałszywą koncepcją umysłu (Fodor 1987: xii). Zgodnie z koncepcjami rozwijanymi przez Fodora (1975, 1987, 1990, 2001) od dziesięcioleci – kognitywistyka doprecyzowuje rzeczywistą naturę postaw propozycjonalnych i rewiduje niektóre przekonania na ich temat. Jednak proponowana przez Fodora, naukowa – przynależąca do pewnego rodzaju obliczeniowej teorii umysłu – taksonomia stanów mentalnych ma być, mówiąc ogólnie, zasadniczo zgodna z psychologią potoczną. Filozof ten twierdzi, że przekonania i pragnienia są wewnętrznymi reprezentacjami – zaimplementowanymi w ośrodkowym układzie nerwowym stanami ob-

liczeniowymi (zdaniami zapisanymi w tak zwanym języku myśli) mającymi pewne treści intencjonalne[17]. Na podstawie rozwijanej przez tego autora koncepcji można uznać, że kiedy wyjaśniamy czyjeś działanie za pomocą pary złożonej z przekonania i pragnienia, odnosimy się do dwóch rzeczywistych stanów (zaimplementowanych w ośrodkowym układzie nerwowym tej osoby), które widnieją również w naukowym, kognitywistycznym wyjaśnieniu tego samego działania. Stany te semantycznie (treściowo, intencjonalnie) i funkcjonalnie odpowiadają stanom psychologii potocznej. Tym samym Fodor jest optymistą, jeśli chodzi o perspektywę uzgodnienia nauk kognitywnych i psychologii potocznej.

Istnieją też jednak zdecydowani filozoficzni pesymiści, jeśli chodzi o możliwość takiej koncyliacji. W głośnym i kontrowersyjnym artykule William Ramsey, Stephen Stich i Joseph Garon (1990) przedstawili argument mający pokazać, że wizja procesów poznawczych zawarta w psychologii potocznej jest nie do pogodzenia z koneksjonizmem, który to stanowi wedle tych autorów źródło najbardziej realistycznych naukowych modeli aktywności systemu poznawczego[18]. Argument przedstawiony na rzecz tej tezy jest dość rozbudowany, jednak na bieżące potrzeby wystarczy pokazać jego zarys. Zgodnie z propozycją Ramseya i współpracowników postawy propozycjonalne mają własność funkcjonalną nazywaną „modularnością propozycjonalną". Własność ta polega na tym, że postawy (1) mogą być indywidualnie nabywane i odrzucane (bez koniecz-

[17] Ujmując rzecz bardziej precyzyjnie, dla Fodora (por. np.: 1990: 17) postawy propozycjonalne są relacjami między podmiotem (organizmem, systemem poznawczym) a wewnętrznymi reprezentacjami. Ujmowanie postaw propozycjonalnych jako relacji między podmiotami a ich własnymi stanami wydaje się dość problematyczne interpretacyjnie. Nie jest oczywiste, jak dokładnie rozumieć ideę, że mając pewne przekonanie, ktoś wchodzi – jako podmiot tego przekonania – w jakąś relację z własnym stanem wewnętrznym (subosobowym). Jednak ostatecznie Fodor wyraża w ten dość specyficzny sposób stosunkowo prostą i zrozumiałą tezę: mieć przekonanie, że *p*, to mieć „w głowie" wewnętrzną, funkcjonalnie scharakteryzowaną reprezentację wyrażającą (reprezentującą) to, że *p*.

[18] A w każdym razie stanowił takie źródło w czasie, kiedy był pisany omawiany artykuł.

ności nabywania lub odrzucania innych przekonań czy pragnień); (2) w sytuacji, gdy osoba jest podmiotem dwóch postaw, z których każda mogła sprawić, że zachowała się ona tak a tak, istnieje definitywna prawdziwa odpowiedź na pytanie o to, która z tych dwóch postaw była rzeczywistą przyczyną tego zachowania. Ramsey i współpracownicy zmierzają do pokazania, że holistyczny i rozproszony sposób przechowywania i przetwarzania informacji w sieciach koneksyjnych wyklucza możliwość, by istniały w tych sieciach stany czy struktury cechujące się modularnością propozycjonalną[19]. Można powiedzieć, że wedle tych autorów znajdujemy się *de facto* w jednej z problematycznych sytuacji wymienionych wcześniej: zachodzi zasadnicza *funkcjonalna* różnica między postawami propozycjonalnymi z jednej strony a stanami czy strukturami postulowanymi przez najlepsze koncepcje z zakresu nauk kognitywnych z drugiej. Wniosek, jaki na tej podstawie formułują Ramsey i współpracownicy, okazuje się druzgocący z punktu widzenia projektu naturalizacji intencjonalności. Przekonania, pragnienia i inne postawy propozycjonalne powinny być całkowicie wyeliminowane z nauk kognitywnych – a szerzej, z naukowego obrazu świata.

Choć poglądy Fodora oraz Ramseya i współpracowników różnią się zasadniczo, to wydaje się, że podzielają oni pewne wspólne założenie. Zgodnie z tym założeniem realizacja ogólnego filozoficznego celu przyświecającego projektowi naturalizacji intencjonalności – celu polegającego na „znalezieniu metafizycznego miejsca" dla

[19] Zauważmy, że strategia tych autorów może być z powodzeniem stosowana nawet w sytuacji, gdy nie dysponujemy teorią opisującą naturalistyczne warunki wystarczające dla bycia postawą propozycjonalną. Ramsey i współpracownicy koncentrują się jedynie na funkcjonalnych własnościach postaw propozycjonalnych, ignorując kwestię ich własności semantycznych. Wystarcza to dla ich teoretycznych celów, ponieważ bycie postawą propozycjonalną wymaga zarówno posiadania określonych własności semantycznych, jak i funkcjonalnych. Wzmiankowani autorzy skupiają się tylko na jednym z warunków koniecznych, lecz niewystarczających do bycia postawą propozycjonalną. Mówiąc zatem ogólnie, przykład Ramseya i współpracowników pokazuje, że można argumentować za eliminacją stanów intencjonalnych pod nieobecność kompletnej teorii specyfikującej wszelkie (naturalistyczne, nieintencjonalne) warunki bycia stanem intencjonalnym.

stanów intencjonalnych w świecie opisywanym za pomocą nieintencjonalnych kategorii przez nauki szczegółowe – wymaga czegoś więcej, niż sformułowania zgodnych z naturalizmem warunków wystarczających dla bycia postawą propozycjonalną. Założenie to można najogólniej scharakteryzować jako ideę, że skuteczna naturalizacja intencjonalności wymaga tego, aby postawy propozycjonalne okazały się przyczynowo aktywnymi stanami wewnętrznymi, które grają rolę w poprawnych kognitywistycznych teoriach i modelach działania systemu poznawczego. Dlatego też powodzenie projektu naturalizacji wymaga w praktyce nie tylko zwycięstwa reprezentacjonizmu w kognitywistyce, ale też zwycięstwa takiej formy reprezentacjonizmu, która postuluje istnienie reprezentacji funkcjonalnie i semantycznie odpowiadających postawom propozycjonalnym.

Ta ważna filozoficzna presupozycja towarzysząca dyskusjom nad możliwością naturalizacji intencjonalności zostanie precyzyjniej wyrażona w dalszej części tej pracy. Na obecnym etapie rozważań chodzi jedynie o zasugerowanie, że relacja między problemem naturalizacji intencjonalności a sporami wokół reprezentacjonizmu prowadzonymi w obrębie kognitywistyki (i filozofii kognitywistyki) *jest sama w sobie ważkim problemem filozoficznym*. Fodor oraz Ramsey i współpracownicy zajmują pewne stanowisko w tej sprawie, jednak nie należy zakładać, iż jest to jedyne stanowisko, które można na ten temat racjonalnie utrzymywać. Niewykluczone, że istnieją filozoficzne podstawy, by ujmować te problemy jako bardziej autonomiczne, niż to przyjmuje wielu współczesnych filozofów. W ostatnich latach pojawiły się głosy sugerujące, że program naturalizacji intencjonalności okazał się fiaskiem (Godfrey-Smith 2004; Lycan 2008). Być może jednak droga naprzód w realizacji tego programu polega nie tyle na podejmowaniu kolejnych prób sformułowania w pełni naturalnych oraz zgodnych z wiedzą naukową wystarczających warunków do bycia przekonaniem czy pragnieniem, lecz w krytycznym przemyśleniu zależności między metafizycznymi „losami" postaw propozycjonalnych i psychologii potocznej a eksplanacyjnymi „losami" pojęcia reprezentacji w naukach kognitywnych.

1.3. W poszukiwaniu koncepcji wyjaśniania reprezentacyjnego

Dotychczasowe ustalenia pozwalają naszkicować mapę przedstawiającą krajobraz dyskutowanych na styku filozofii i kognitywistyki problemów związanych z pojęciem reprezentacji mentalnych. Powinny zostać rozróżnione następujące „problemy reprezentacji":

1. Problem statusu eksplanacyjnego reprezentacji mentalnych w kognitywistyce:

a) przedmiotowy: czy kognitywiści potrzebują pojęcia reprezentacji do realizacji stawianych sobie celów eksplanacyjnych? Czy dobre (poprawne, najlepsze spośród dostępnych) wyjaśnienia odwołują się do wewnętrznych reprezentacji? Czy (lub w jakim zakresie) system poznawczy jest systemem reprezentacyjnym?

b) metaprzedmiotowy: na czym polega wyjaśnienie danego zjawiska za pomocą reprezentacji? Jak odróżnić pełnoprawnie reprezentacyjne wyjaśnienie od takiego, które nie jest reprezentacyjne lub jest reprezentacyjne jedynie pozornie? Na jakiej podstawie możemy stwierdzić, że badany system rzeczywiście posługuje się reprezentacjami?

2. Problem naturalizacji intencjonalności: jak stany intencjonalne (postawy propozycjonalne) mogą być identyczne ze stanami naturalistycznie pojętego systemu poznawczego? Jakie są czysto naturalne własności, które może egzemplifikować wewnętrzny stan systemu poznawczego, a których posiadanie wystarcza do tego, by stan ten był identyczny z postawą propozycjonalną o określonej treści?

3. Problem relacji między projektem naturalizacji intencjonalności a problemem statusu eksplanacyjnego reprezentacji w kognitywistyce: czy powodzenie projektu naturalizacji intencjonalności zależy od faktów dotyczących eksplanacyjnej wartości reprezentacji dla kognitywistów? Czy powodzenie projektu naturalizacji zależy od tego, jakie własności intencjonalne i funkcjonalne przysługują reprezentacjom postulowanym przez

kognitywistów (zakładając, że reprezentacje są w ogóle przez nich postulowane)? Czy psychologia potoczna wymaga naukowej „legitymizacji"?

Powyższe zestawienie rzecz jasna mogłoby być dużo bardziej szczegółowe. Zaznaczyłem wcześniej na przykład, że problem naturalizacji intencjonalności generuje cały szereg bardziej szczegółowych wyzwań. Z kolei w (przedmiotowym) problemie statusu eksplanacyjnego reprezentacji w kognitywistyce należałoby odróżnić rozwiązania „globalne" (dotyczące systemu poznawczego jako takiego) od „lokalnych" (zrelatywizowanych do konkretnych eksplanandów). Mimo to, powyżej przedstawiona lista jest wystarczająco szczegółowa, aby można było na jej podstawie precyzyjnie wyznaczyć teoretyczne cele, jakie będą realizowane w kolejnych rozdziałach.

Głównym problemem, jaki zostanie tu podjęty, jest *metaprzedmiotowy* problem statusu eksplanacyjnego reprezentacji mentalnych w kognitywistyce. Podstawowym celem teoretycznym tej książki będzie zatem sformułowanie oraz uzasadnienie określonej koncepcji wyjaśniania reprezentacyjnego w naukach kognitywnych. Przedstawione tu rozważania mają odpowiedzieć na pytanie, czym są wyjaśnienia tego rodzaju. Mówiąc inaczej, moim nadrzędnym celem jest pokazanie, jakie warunki powinno spełniać kognitywistyczne wyjaśnienie danego zjawiska, abyśmy mogli w uprawniony sposób uznać je za wyjaśnienie reprezentacyjne.

Na podstawie poczynionych w tym rozdziale ustaleń powinno być jasne, z jak ważnym i niecierpiącym zwłoki zagadnieniem mamy do czynienia. Istnieją trzy zasadnicze powody, dla których metaprzedmiotowy problem reprezentacji należy uznać za ważny i „pilny". Po pierwsze, rozwiązanie metaprzedmiotowego sporu o reprezentacjonizm to warunek *sine qua non* rozwiązania tego sporu w jego przedmiotowym wymiarze. Ta ostatnia kwestia ma fundamentalne znaczenie dla nauk kognitywnych i jest szczególnie istotna w momencie, gdy antyreprezentacjonizm zyskuje do tej pory niespotykane poparcie, przedostając się z peryferii kognitywistyki do jej głównego nurtu. Dopóki metaprzedmiotowy problem reprezentacji nie zostanie rozwiązany, dopóty przedmiotowy problem repre-

zentacjonizmu jest toczony, przynajmniej częściowo, „po omacku" (por. Haselager, de Groot, van Rappard 2003). Po drugie, w literaturze brakuje systematycznych opracowań metaprzedmiotowego problemu eksplanacyjnego statusu reprezentacji mentalnych (por. jednak istotne wyjątki: Ramsey 2007; Miłkowski 2013). Książka ta dotyczy zatem zagadnienia stosunkowo „niedocenionego" w literaturze; liczba podjętych do tej pory prób rozwiązania, a nawet sformułowania podejmowanego tu problemu wydaje się nieproporcjonalna do jego wagi. Po trzecie, założenia o tym, czym są wyjaśnienia reprezentacyjne, są na ogół przemycane *implicite* i często nie są podzielane przez obie strony sporu między reprezentacjonistami a antyreprezentacjonistami. Wyniki prowadzonych tu rozważań powinny zatem umożliwić wprowadzenie jasności i porządku w obszarze problemowym, w którym obecnie panuje pojęciowe zamieszanie.

Rozwiązanie metaprzedmiotowego problemu reprezentacjonizmu wymaga jednak sformułowania określonych kryteriów, których spełnienie gwarantuje danemu wyjaśnieniu status wyjaśnienia reprezentacyjnego. Należy więc oczekiwać koncepcji, która pozwoli w możliwie precyzyjny sposób odróżnić wyjaśnienia reprezentacyjne od niereprezentacyjnych. Ponadto koncepcja ta powinna umożliwić odróżnienie wyjaśnień zasadnie reprezentacyjne od tych, które są takimi jedynie deklaratywnie, bo posługują się pojęciem reprezentacji na tyle liberalnie, że przypisują ten status stanom czy strukturom, które na niego nie zasługują (por. Ramsey 2007: 1–37). Jak jednak sformułować kryterium lub zestaw kryteriów pozwalających na poczynienie tych rozróżnień? Wydaje się, że przydatne w tym celu może być postawienie prowadzonych tu rozważań w kontekście szerszej koncepcji *wyjaśniania naukowego*, koncepcji, która może być zaaplikowana do zagadnienia natury wyjaśniania w kognitywistyce. Odpowiedź na pytanie o to, na czym polega wyjaśnianie w kognitywistyce jako takie powinna pomóc w udzieleniu odpowiedzi na pytanie o to, czego dokładnie powinniśmy oczekiwać od wyjaśnień reprezentacyjnych. Dlatego też będę w tej pracy przyjmował, że kryterium odróżniania wyjaśnień reprezentacyjnych od niereprezentacyjnych należy poszukiwać, posiłkując się osiągnięciami filozofii nauki.

W ten właśnie sposób dochodzimy do założenia, które odegra zasadniczą rolę w toku prowadzonych tu rozważań. Głosi ono, że wyjaśnienia w kognitywistyce przyjmują na ogół formę *wyjaśnień mechanistycznych*. Wyjaśnianie zjawisk przez kognitywistów opiera się na odkrywaniu i opisywaniu stojących u ich podstaw mechanizmów. Jeśli zaaplikować tę ogólną ideę do głównego celu teoretycznego tej pracy, można powiedzieć, że kryteria bycia reprezentacyjnym wyjaśnieniem w kognitywistyce będą wyznaczane przez kryteria bycia reprezentacyjnym wyjaśnieniem mechanistycznym. Jak zobaczymy dalej, oznacza to, że ostatecznym celem prowadzonych tu rozważań jest stworzenie ogólnej koncepcji mechanizmów reprezentacyjnych – mechanizmów, których działanie opiera się na wewnętrznych reprezentacjach. To ostatnie sformułowanie dotyczy już nie tyle wyjaśnień, co raczej natury samych mechanizmów. Jednakże pytanie o naturę mechanizmów reprezentacyjnych będzie w tej książce traktowane jako równoważne pytaniu o naturę mechanistycznych wyjaśnień reprezentacyjnych. Jest to zgodne z zaznaczoną tu już kilkakrotnie ideą, że epistemiczne kryteria bycia wyjaśnieniem reprezentacyjnym są ściśle związane z metafizycznymi kryteriami bycia systemem reprezentacyjnym[20]. W dalszym toku rozważań pokażę, że taki sposób myślenia naturalnie wynika z mechanistycznej koncepcji wyjaśniania naukowego.

Główną bronioną na kartach tej książki tezą jest propozycja, że mechanizmy reprezentacyjne to mechanizmy, których działanie opiera się na *konsumowanych modelach*. Oznacza to po pierwsze, że mechanizm reprezentacyjny działa na podstawie wewnętrznego modelu. Może to być model środowiska zewnętrznego lub ciała własnego. Bycie modelem jest (współ)konstytuowane przez fakt, że nośnik reprezentacji pozostaje *strukturalnie podobny* do tego, co reprezentowane. Po drugie, mechanizm reprezentacyjny to taki mechanizm, w którym wewnętrzny model *pełni funkcję modelu*. Dochodzi do tego wtedy, gdy w ramach mechanizmu istnieje komponent „konsumujący" reprezentację, wykorzystujący ją w celu poprawnego re-

[20] Przez „bycie systemem reprezentacyjnym" będę rozumieć w tej pracy bycie systemem działającym na podstawie mechanizmów reprezentacyjnych.

alizowania pełnionej przez siebie funkcji[21]. Parafrazując tę tezę, tak aby dotyczyła ona wyjaśniania (a nie samych mechanizmów), można powiedzieć, że wyjaśnienie reprezentacyjne danego zjawiska to wyjaśnienie tego zjawiska za pomocą mechanizmu wyposażonego w konsumowany model pewnej domeny.

Sformułowanie kryteriów, za pomocą których można ocenić, kiedy reprezentacje mentalne spełniają w sposób zasadny swoją rolę eksplanacyjną, pozwala także na rozpoznanie sytuacji, w których tak nie jest. Dotyczy to nie tylko wyjaśnień określanych w otwarty sposób jako antyreprezentacjonistyczne. Chodzi także o możliwość rozpoznania sytuacji, w których wyjaśnienia *uznawane* za reprezentacyjne w rzeczywistości takimi nie są. Nie znaczy to z konieczności, że wyjaśnienia te są niepoprawne. Rzecz raczej w tym, że postulowane w nich stany czy struktury nie spełniają swoich ról eksplanacyjnych *jako reprezentacje*. W toku prowadzonego tu wywodu będę wielokrotnie nawiązywać do rozważań Ramseya (2007), który argumentuje, że spora część wyjaśnień zjawisk poznawczych formułowanych przez kognitywistów na przestrzeni ostatnich dziesięcioleci to wyjaśnienia jedynie pozornie czy nominalnie reprezentacyjne[22].

Podsumowując dotychczasowe ustalenia: głównym celem tej książki jest zaproponowanie rozwiązania metaprzedmiotowego problemu statusu eksplanacyjnego reprezentacji mentalnych w kognitywistyce. Zgodnie z główną bronioną tu tezą wyjaśnienia reprezentacyjne to wyjaśnienia odwołujące się do mechanizmów wykorzystujących konsumowane modele. Teoretycznym tłem dla moich rozważań będzie zaś wywodząca się z filozofii nauki koncepcja wyjaśniania mechanistycznego.

---

[21] Pojęcie „konsumenta reprezentacji" zapożyczam z koncepcji Ruth Millikan (1984, 2002), jednak z istotnymi modyfikacjami. Różnice między rozumieniem konsumentów reprezentacji w teorii Millikan i teorii bronionej w mojej pracy zostaną omówione w sekcji 4.2.1.

[22] Jak zobaczymy, niektóre idee sformułowane przez Ramseya w jego pracy *Representation Reconsidered* (2007) odegrają znaczącą rolę w prowadzonym tu wywodzie. Prezentowana książka stanowi do pewnego stopnia próbę uzupełnienia i rozwinięcia niektórych twierdzeń wyrażonych przez tego autora, jak również próbę podjęcia polemiki z niektórymi z nich (zwłaszcza z argumentacją Ramseya na rzecz antyreprezentacjonizmu – por. podrozdział 4.3).

Traktowanie wymienionych na początku tego podrozdziału „problemów reprezentacji" jako całkowicie niezależnych byłoby pomyłką. Między tymi zagadnieniami zachodzą bardziej lub mniej widoczne na pierwszy rzut oka interakcje. Można zatem oczekiwać, że poczynione tu ustalenia dotyczące metaprzedmiotowego problemu statusu eksplanacyjnego reprezentacji pozwolą na wyciągnięcie określonych konsekwencji zarówno dla *przedmiotowego* problemu reprezentacji w kognitywistyce, jak i dla zagadnień związanych z projektem *naturalizacji intencjonalności*. Wskazanie i opisanie tych konsekwencji wyznacza dwa kolejne cele teoretyczne tej książki.

Drugim celem moich rozważań będzie zajęcie stanowiska w sprawie przedmiotowego problemu statusu eksplanacyjnego reprezentacji mentalnych w kognitywistyce. Będę zmierzał do pokazania, że bronione często twierdzenia o schyłku kognitywistyki opartej na pojęciu reprezentacji mentalnych są nieuzasadnione i przedwczesne. Zgodnie z zajmowaną tu przeze mnie (na poziomie metaprzedmiotowym) pozycją wyjaśnienia mechanistyczne odwołujące się do konsumowanych modeli są pełnoprawnie reprezentacyjne. Dopóki konsumowane modele stanowią istotny element eksplanacyjnego repozytorium nauk kognitywnych, dopóty poprawne pozostaje twierdzenie, że reprezentacje stanowią istotny element tego repozytorium. Mówiąc obrazowo, proponowane tu rozwiązanie metaprzedmiotowego problemu statusu eksplanacyjnego reprezentacji pozwala wyznaczyć warunki, w których kognitywistyka jest (na poziomie przedmiotowym) reprezentacjonistyczna. W ramach tej pracy zostanie wskazany szereg przykładowych, teoretycznie i empirycznie uzasadnionych wyjaśnień zjawisk kognitywnych, które odwołują się do mechanizmów wyposażonych w konsumowane modele, a zatem spełniają sformułowane przeze mnie kryteria bycia wyjaśnieniem reprezentacyjnym. Koncepcje te są nowe lub stosunkowo nowe w teoretycznym krajobrazie kognitywistyki: powstały na przestrzeni ostatnich dwudziestu-trzydziestu lat, już po tym, jak szczytową popularność osiągnął koneksjonizm. Nie należą one zatem do przeszłości nauk kognitywnych, lecz stanowią ich teraźniejszość. Tym samym można twierdzić, że – wbrew głosom antyreprezentacjonistów – kognitywistyka nie znajduje się w postrepre-

zentacjonistycznym stadium swojej historii. Istnieją podstawy, by sądzić, że system poznawczy jest, przynajmniej w pewnym nietrywialnym stopniu, systemem reprezentacyjnym.

Trzecim wreszcie celem stawianym w tej pracy będzie wskazanie konsekwencji, jakie przyjęte tu mechanistyczne podejście do wyjaśniania reprezentacyjnego w kognitywistyce niesie dla projektu naturalizacji intencjonalności. Mówiąc dokładniej, skupię się na przemyśleniu relacji między projektem naturalizacji a kwestią eksplanacyjnego statusu reprezentacji w kognitywistyce. Kluczową rolę w mojej propozycji odegra fakt, że podejście mechanistyczne pozwala na nadanie precyzyjnego sensu idei, że systemy poznawcze są systemami *wielopoziomowymi*. Cechują się one wielopoziomową organizacją oraz mogą być na różnych poziomach analizowane (opisywane i wyjaśniane). Postaram się pokazać, że wyjaśnienia działań formułowane za pomocą psychologii potocznej – a zatem te wyjaśnienia, które odwołują się do postaw propozycjonalnych – znajdują się na innym poziomie niż wyjaśnienia formułowane w ramach kognitywistyki. Psychologia potoczna może zostać zinterpretowana jako *mechanistycznie neutralne* narzędzie eksplanacyjne. Jeśli jest to poprawna teza, to okazuje się, że powodzenie projektu naturalizacji intencjonalności nie wymaga z konieczności, aby było możliwe zidentyfikowanie postaw propozycjonalnych z wewnętrznymi stanami lub strukturami systemu poznawczego (czyli, ostatecznie, wewnętrznymi stanami lub strukturami biologicznego mózgu). Oznacza to, że postawy propozycjonalne mogą być realne nawet wtedy, gdy mechanistyczna dekompozycja systemu poznawczego pokaże, iż na poziomie jego wewnętrznej organizacji nie istnieje nic, co funkcjonalnie lub intencjonalnie tym postawom odpowiada. Spróbuję zatem pokazać, że filozoficzne „wymagania" nakładane zazwyczaj na udaną naturalizację intencjonalności są zbyt restrykcyjne. Problem statusu eksplanacyjnego reprezentacji w kognitywistyce i problem naturalizacji intencjonalności są bardziej autonomiczne, niż się na ogół zakłada.

Jak zatem widać, chociaż prezentowana książka koncentruje się na metaprzedmiotowym aspekcie sporu między reprezentacjonistami a antyreprezentacjonistami, to jej główne założenia i tezy niosą ze

sobą konsekwencje obejmujące sporą część krajobrazu problemowego związanego z pojęciem reprezentacji mentalnych. Teoretycznym spoiwem łączącym wszystkie bronione tu idee pozostaje jednak założenie, że system poznawczy to hierarchiczny układ mechanizmów, a wyjaśnianie w naukach kognitywnych stanowi formę wyjaśniania mechanistycznego. Zanim będzie można przejść do realizacji postawionych tu celów teoretycznych, należy zatem przedstawić mechanistyczne podejście do kwestii wyjaśniania w kognitywistyce.

Rozdział 2

Mechanistyczny model wyjaśniania naukowego

2.1. Mechanicyzm: ogólna charakterystyka

2.1.1. Wyjaśnianie mechanistyczne. Natura, badanie i naukowe reprezentowanie mechanizmów

Na wstępie należy zaznaczyć, że przedstawione tu omówienie koncepcji wyjaśniania mechanistycznego nie będzie aspirować do miana kompletnej rekonstrukcji. Pominę tu różne punkty sporne, które odróżniają poszczególne sformułowania mechanicyzmu. Moja rekonstrukcja będzie prowadzona tak, by wyodrębnić „rdzeń" koncepcji wyjaśniania mechanistycznego, to znaczy te twierdzenia i założenia, co do których wśród poszczególnych zwolenników współczesnego mechanicyzmu istnieje znaczny konsensus. Co więcej, literatura dotycząca mechanicyzmu oraz poszczególnych metafizycznych, epistemologicznych i metodologicznych zagadnień powstających w jego kontekście jest w tej chwili na tyle szeroka, że jej pełne omówienie nie wydaje się ani niezbędne, ani wskazane z punktu widzenia celów tej książki. Przedstawiona tu rekonstrukcja będzie więc selektywna. Pewne wątki – chociażby problem relacji między mechanicyzmem a kwestią jedności nauki, zmianą naukową czy naturą przyczynowości – zostaną pominięte. Pominę też kwestię historycznych źródeł i inspiracji mechanistycznego modelu wyjaśniania. Zostanie on tu omówiony w postaci, jaką przyjął na przełomie XX i XXI wieku, głównie w pracach Williama Bechtela i Carla Cravera (prace, na których opiera się przedstawiona tu rekonstrukcja, to w szczególności: Bechtel, Richardson 1993; Cummins 2000; Glennan 2002; Bechtel, Abrahamsen 2005; Craver 2007; Bechtel 2008; Machamer, Darden, Craver 2011). Na czym zatem skupię się w tym rozdziale? Najpierw omówię ogólne założenia i najważniejsze tezy mechanistycznego modelu wyjaśniania. Podejmę też kwestię związku mechanicyzmu

z alternatywnymi modelami wyjaśniania, w szczególności zaś zagadnienie, w jaki sposób mechanicyzm odbiega od szeregu klasycznych założeń dotyczących wyjaśniania opartych na modelu nomologiczno-dedukcyjnym. W kolejnych częściach rozdziału przyjrzę się bliżej tym zagadnieniom związanym z mechanicyzmem, które są bezpośrednio istotne dla wywodu prowadzonego w dalszej części tej pracy. Będą to: (1) zagadnienie relacji międzypoziomowych w wyjaśnieniach mechanistycznych (podrozdział 2.2); (2) kwestia aplikowalności mechanicyzmu do zagadnienia natury wyjaśniania w naukach kognitywnych (podrozdział 2.3). Dyskusję warto jednak zacząć od przedstawienia szeregu podstawowych założeń i tez mechanicyzmu.

U podstaw koncepcji wyjaśniania mechanistycznego stoi twierdzenie, że naukowcy są często zaangażowani w wyjaśnianie działania złożonych systemów w przyrodzie, a osiągają ten cel, odkrywając mechanizmy odpowiedzialne za to działanie. Praktykę poznawczą badaczy motywuje często pytanie: „jak to działa?", a udzielenie na nie odpowiedzi wymaga opisania stosownego mechanizmu. Taka ogólna teza nie jest rzecz jasna odkrywcza. Wydaje się, że naukowcy mówią często o mechanizmach i bardziej lub mniej otwarcie konceptualizują własną praktykę eksplanacyjną w kategoriach odkrywania i opisywania mechanizmów. Wartość rozwijanych współcześnie koncepcji wyjaśniania mechanistycznego polega jednak przede wszystkim na dostarczeniu precyzyjnej i bogatej w filozoficzne konsekwencje analizy tego, czym dokładnie są mechanizmy oraz tego, na czym polegają różnice między perspektywą mechanistyczną a pewnymi zakorzenionymi w tradycji filozoficznej założeniami dotyczącymi natury wyjaśniania naukowego. Zajmijmy się najpierw kwestią zasadniczą, dotyczącą tego, czym jest mechanizm, oraz co to znaczy, że wyjaśnia on pewne zjawisko. Nie istnieje co prawda całkowity konsensus dotyczący tego, jak dokładnie należy definiować (czy scharakteryzować) mechanizmy. Mimo to ogólna idea została zawarta w poniższych, zaczerpniętych z literatury propozycjach:

> Mechanizm to struktura pełniąca pewną funkcję dzięki swoim komponentom, operacjom komponentów oraz ich organizacji. Skoordynowane funkcjonowanie mechanizmu jest odpowiedzialne za

> powstanie jednego lub wielu różnych zjawisk (Bechtel, Abrahamsen 2005: 423).
>
> Mechanizmy to jednostki (*entities*) oraz działania (*activities*) zorganizowane, tak że razem egzemplifikują one zjawisko stanowiące eksplanandum (Craver 2007: 6).

Niektóre elementy powyższych definicji nie są całkowicie powszechnie akceptowane w literaturze dotyczącej mechanicyzmu[1]. Pomijając jednak pewne subtelne różnice między poszczególnymi koncepcjami, w ramach tej pracy będę przyjmować rozumienie mechanizmów jako zorganizowanych układów komponentów i realizowanych przez te komponenty operacji. Mechanizmy zatem:

a) to układy czy całości złożone z elementów składowych, czyli *komponentów*,

b) ich komponenty są zaangażowane w różnego rodzaju *operacje* (działania, czynności), to znaczy realizują one w ramach mechanizmu określone funkcje,

c) ich komponenty oraz wykonywane przez nie działania charakteryzują się określoną przestrzenną, temporalną i przyczynową *organizacją*,

d) są hierarchicznie *wielopoziomowe*[2],

e) stoją u podstaw czy też odpowiadają za określone *zjawisko* (stanowiące eksplanandum wyjaśnienia mechanistycznego).

---

[1] Na przykład Stuart Glennan (2002) w swojej charakterystyce mechanizmów nie wyróżnia operacji jako kategorii osobnej w stosunku do komponentów (jednostek). Autor ten postuluje, że mechanizmy zawierają jedynie komponenty, między którymi zachodzą określone interakcje. Warto mieć jednak na uwadze, że wcale nie jest pewne, czy między stanowiskiem Glennana (postulującym interreagujące komponenty) a alternatywą (postulującą komponenty oraz wykonywane przez nie działania czy operacje) zachodzi teoretycznie istotna różnica (Tabery 2004). Operacje przypisywane komponentom polegają często na tym, że dany komponent ma określony wpływ na jakiś inny komponent. Innymi słowy, przypisywanie komponentom operacji to *de facto* charakteryzowanie ich interakcji z innymi komponentami.

[2] Wielopoziomowość nie została co prawda wymieniona w przytoczonych wyżej definicjach, jednak stanowi ważną oraz powszechnie akceptowaną własność mechanizmów.

Przyjrzyjmy się bliżej poszczególnym elementom tej charakterystyki.

Ad a) Komponenty

Mechanizmy mają naturę kompozycyjną: są one układami zorganizowanych części czy komponentów. Mechanistyczne wyjaśnienie zdolności organizmu do dystrybucji tlenu i składników odżywczych do komórek jest oparte na wskazaniu mechanizmu, w którego obręb wchodzą serce, krew, żyły, tętnice i naczynia włosowate. Są to komponenty tego mechanizmu. Charakterystykę czegoś w kategoriach składających się na nie komponentów nazywa się „dekompozycją strukturalną" (Bechtel, Richardson 1993; Bechtel 2008).

Trzeba wyraźnie zaznaczyć, że nie każda możliwa dekompozycja strukturalna badanego systemu fizycznego będzie skuteczna z perspektywy zadania polegającego na dostarczeniu mechanistycznego wyjaśnienia cechujących ten system zdolności. Na przykład jeśli jesteśmy zainteresowani odkryciem mechanizmów neuronalnych odpowiadających za zdolności poznawcze, to nie uda nam się tego dokonać na podstawie strukturalnej dekompozycji mózgu według bruzd, w które układa się kora, jak również na podstawie dekompozycji odwołującej się do cytoarchitektury (organizacji komórkowej) odpowiednich obszarów mózgu (Bechtel 2008: 15–17). Dzieje się tak dlatego, że oba te podziały na komponenty nie wyróżniają (lub wyróżniają jedynie bardzo nieprecyzyjnie) komponentów *aktywnych*, czyli takich, które są rzeczywiście zaangażowane funkcjonalnie w umożliwianie wyjaśnianego zjawiska. Innymi słowy, jednostkom czy komponentom wyróżnionym na podstawie układu bruzd lub na podstawie cytoarchitektury nie odpowiadają określone operacje, czynności czy role funkcjonalne w ramach mechanizmów odpowiadających za zjawiska poznawcze. Tymczasem to właśnie takie aktywne komponenty – które „robią coś" w ramach mechanizmu odpowiedzialnego za dane zjawisko – są istotne w projekcie dostarczania wyjaśnień mechanistycznych. Z tego powodu wspomniana powyżej lista komponentów układu krwionośnego (jako mechanizmu odpowiedzialnego za dostarczenie tlenu i składników odżywczych do komórek) zawiera takie komponenty, których operacje rzeczywiście

stoją u podstaw zjawiska stanowiącego eksplanandum. Jeden z tych komponentów funkcjonuje w ramach mechanizmu jako nośnik tlenu i składników odżywczych, inny jako pompa umożliwiająca przepływ tego nośnika, a jeszcze inne komponenty służą jako „szlaki", za których pośrednictwem nośnik porusza się w obrębie organizmu. Poszukując zatem komponentów mechanizmu, poszukujemy takich komponentów, których czynności są funkcjonalnie relewantne dla zjawiska przez nas wyjaśnianego. W perspektywie mechanistycznej dekompozycja strukturalna nie jest niezależna od dekompozycji *funkcjonalnej*, czyli dekompozycji czegoś na operacje składowe (Bechtel, Richardson 1993; Craver 2007; Bechtel 2008; Piccinini, Craver 2011).

Ad b) Operacje

Powyższa konstatacja naturalnie kieruje nas ku zagadnieniu *operacji* wykonywanych przez komponenty. Ujmując to niezbyt technicznie, operacja wykonywana przez dany komponent jest tym, co ten komponent „robi" w ramach mechanizmu. Operacjami, jakie są przypisywane przez naukowców komponentom różnych mechanizmów, mogą być choćby „pompowanie", „filtrowanie", „katalizowanie", „przenoszenie" czy „wydzielanie"[3]. Przykłady operacji przypisywanych poszczególnym komponentom układu krwionośnego zostały już wymienione w punkcie (a).

Warto podjąć zagadnienie relacji między operacjami komponentów mechanizmu a pełnionymi przez nie *funkcjami* (Craver 2001, 2008; Machamer, Darden, Craver 2011). Z jednej strony można scharakteryzować operacje niezależnie od ich wkładu w funkcjonowanie pewnego szerszego mechanizmu. Operacje w tym sensie mogą być dowolnymi, scharakteryzowanym bezkontekstowo wzorami lub regularnościami w zachowaniu komponentu, czy też korelacjami zachodzącymi między jego stanem początkowym a końcowym. Z takiego punktu widzenia sercu można przypisać zarówno

[3] Jak zobaczymy w dalszej części tej pracy, z punktu widzenia mechanicyzmu kluczem do zagadnienia eksplanacyjnej roli reprezentacji w naukach kognitywnych jest właśnie pytanie o to, w jakim sensie reprezentowanie może być operacją wykonywaną przez komponent mechanizmu.

operację polegającą na pompowaniu krwi (zdefiniowaną jako korelacja między stanami, w jakich znajduje się serce w jakimś punkcie „początkowym" i „końcowym"), jak i operację polegającą na emitowaniu rytmicznych dźwięków (Cummins 1975). Z drugiej jednak strony z punktu widzenia eksplanacyjnych zainteresowań nauki istotne są operacje, które mają charakter funkcjonalny dla działania mechanizmu jako całości (Craver 2001, 2008; Machamer, Darden, Craver 2011). Funkcjonalność operacji polega na tym, że jest ona istotna dla poprawnego funkcjonowania mechanizmu, to znaczy dla tego, że mechanizm ten umożliwia zachodzenie danego zjawiska. W takiej perspektywie przypisanie operacji wymaga weryfikacji, jak dany komponent pozostaje przyczynowo (interakcyjnie) usytuowany w ramach szerszego mechanizmu oraz w jaki sposób wpływa on (w ramach tego kontekstu) na zjawisko wyjaśniane przez ten mechanizm. W ramach mechanistycznego wyjaśnienia dystrybucji tlenu i składników odżywczych do tkanek organizmu, operacja serca polega na pompowaniu krwi, ale już nie na generowaniu dźwięków. Uogólniając, operacja w węższym sensie to zawsze czynność, która stanowi *funkcję* komponentu w ramach mechanizmu. W dalszej części tej pracy będę posługiwać się właśnie takim rozumieniem operacji komponentów.

Powyższe stwierdzenia na temat relacji między operacjami a funkcjami „przemycają" rzecz jasna pewnego rodzaju teorię funkcji. Chodzi mianowicie o teorię funkcji jako ról przyczynowych, wywiedzioną z klasycznej pracy Roberta Cumminsa (1975; por. też Craver 2001, 2008). Zgodnie z tą koncepcją:

> Komponent x służy jako φ w ramach S (x pełni funkcję φ w ramach S) relatywnie do mechanistycznego wyjaśnienia M zdolności S do ψ, wtedy, gdy x realizuje działanie φ w ramach M oraz M poprawnie wyjaśnia zdolność S do wykonywania ψ przez, między innymi, odwołanie się do tego, że x wykonuje działanie φ w ramach S (jest to zmodyfikowana wersja propozycji przedstawionej w: Cummins 1975: 762).

O realizowaniu określonej funkcji decyduje zatem osadzenie w szerszym mechanizmie odpowiadającym za określone zjawisko. Na przykład o pełnieniu przez serce funkcji pompy decyduje fakt, że to jego operacja jako pompy jest istotna dla poprawnego funkcjonowania mechanizmu odpowiedzialnego za transport tlenu i składników odżywczych w obrębie organizmu. Innymi słowy, funkcję komponentu wyznacza ta operacja, która jest mu przypisywana w ramach (poprawnego) mechanistycznego wyjaśnienia danego zjawiska; operacja, której wykonywanie przez ten komponent jest relewantne dla zjawiska, jakie wyjaśniamy za pomocą określonego mechanizmu. Takie ujęcie funkcji różni się choćby od teorii etiologicznej, gdzie funkcjonalny status „zawdzięcza" się posiadaniu odpowiedniej historii ewolucyjnej. Funkcje w sensie Cumminsa nie są funkcjami właściwymi w sensie Ruth Millikan (1984). Z punktu widzenia tego pierwszego autora, przypisywanie ról funkcjonalnych komponentom mechanizmów nie niesie żadnych zobowiązań dotyczących historii presji selekcyjnych odpowiadających za powstanie i gatunkowe rozpowszechnienie tych komponentów. Kiedy zatem przypisujemy komponentom mechanizmu wykonywanie w nim określonych operacji, mamy na myśli, że komponenty te realizują w tym mechanizmie (względem innych komponentów) określone *funkcje w sensie Cumminsa*.

Ad c) Organizacja

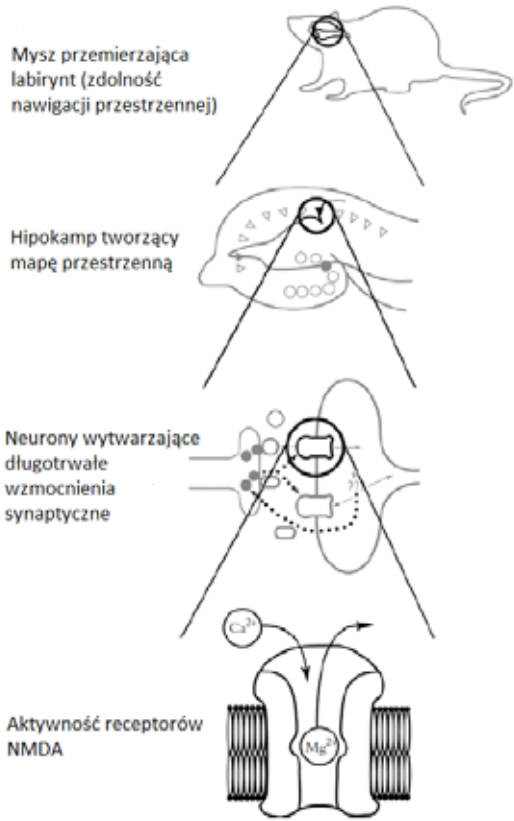
Mechanizmy nie są prostymi agregatami, lecz charakteryzują się zawsze pewnego rodzaju organizacją[4]. Tylko odpowiednio zorganizowane układy komponentów i operacji mogą być odpowiedzialne za zjawiska stanowiące eksplananda wyjaśnień mechanistycznych. Craver (2007: 136–139) wyróżnia trzy aspekty organizacji mechanizmów: przestrzenny, temporalny oraz aktywny/przyczynowy (*active*). Organizacja przestrzenna wiąże się ze wzajemnym usytuowaniem komponentów, ich kształtami i wielkością czy kierunkiem ruchu. Organizacja temporalna dotyczy między innymi porządku

[4] Zagadnienie nieagregatywnej natury mechanizmów zostanie jeszcze omówione w dalszej części tego rozdziału (w sekcji 2.2.2).

czasowego, czasu trwania czy częstotliwości operacji wykonywanych przez komponenty. Wreszcie organizacja aktywna jest związana ze wzorem czy strukturą przyczynowych interakcji zachodzących między komponentami mechanizmu – ze sposobem, w jaki ich działania są wzajemnie od siebie zależne. Jak stwierdza Craver, „mechanizmy [...] nie są jedynie statycznymi czy przestrzennymi wzorami relacji, lecz raczej wzorami umożliwiania, generowania, blokowania, produkowania i stymulowania" (2007: 136). Rzecz jasna odróżnienie tych trzech aspektów organizacji mechanizmów to zabieg czysto analityczny. W rzeczywistych mechanizmach wszystkie one są ze sobą ściśle powiązane, a określona organizacja aktywna nie jest możliwa, jeśli nie utrzymuje się ona w ramach określonej organizacji przestrzennej i temporalnej. Zachowanie mechanizmu zależy od tego, jakie wzory interakcji zachodzą między jego komponentami i w jakim porządku przestrzennym oraz temporalnym one zachodzą.

Ad d) Wielopoziomowość

Kolejną ważną cechę mechanizmów stanowi ich wielopoziomowość. Wiąże się ona z ich hierarchicznie kompozycyjną organizacją. Operacja realizowana przez komponent danego mechanizmu może stanowić eksplanandum i zostać wyjaśniona przez mechanizm znajdujący się na niższym poziomie organizacji. Na przykład pompowanie krwi przez serce w ramach mechanizmu odpowiadającego za dystrybucję tlenu i składników odżywczych do komórek organizmu może być potraktowane jako eksplanandum – jako zdolność, którą można wyjaśnić mechanistycznie. Takie wyjaśnienie będzie się odwoływać do mechanizmu znajdującego się na niższym poziome organizacji, którego działanie jest oparte (w uproszczeniu) na zorganizowanym układzie skurczających i rozkurczających się komór. Weźmy też pod uwagę inną ilustrację, zaczerpniętą z neuronauki. Bechtel (2009) i Craver (2007: 164–170) omawiają przykład neuronaukowego, hierarchicznie mechanistycznego wyjaśnienia pamięci przestrzennej u myszy i szczurów (por. rysunek 2). Na najwyższym poziomie – czyli na poziomie systemu poznawczego lub organizmu jako całości – mamy w tym przypadku do czynienia ze zdolnością



Rysunek 2. Hierarchia mechanizmów odpowiadających za zdolność nawigacji przestrzennej u szczurów. Źródło: Craver 2007: 166

do pamiętania swojego otoczenia przestrzennego. Dostarczenie dokładnego opisu tej kompetencji stanowi samo w sobie wyzwanie wymagające badań eksperymentalnych koncentrujących się na systemie (organizmie) jako całości oraz jego interakcjach z otoczeniem. Mechanistyczne wyjaśnienie pamięci przestrzennej wymaga jednak zejścia na niższy poziom organizacji systemu poznawczego. W tym przypadku stosowny mechanizm opiera się najprawdopodobniej na

tworzeniu wewnętrznych map przestrzennych środowiska w hipokampie oraz skroniowych i czołowych strukturach korowych. Chociaż tworzenie map tego rodzaju stanowi tu eksplanans, to wyznacza ono jednocześnie *eksplanandum* wyjaśnienia odwołującego się do mechanizmu z kolejnego (niższego) poziomu. Chodzi mianowicie o mechanizm długotrwałego wzmocnienia synaptycznego neuronów hipokampu, który to sam może zostać wyjaśniony za pomocą mechanizmu z jeszcze niższego (molekularnego) poziomu organizacji, czyli przez aktywność receptorów NMDA. Jak zatem widzimy, dostarczenie mechanistycznego wyjaśnienia zdolności orientacji przestrzennej wymagało odkrycia i opisania „szkatułkowego" układu mechanizmów znajdujących się na coraz niższych „stopniach" hierarchicznej organizacji – począwszy od poziomu organizmu (systemu poznawczego) jako całości wchodzącej w interakcje ze środowiskiem, aż po poziom molekularny.

Ad e) Zjawiska (i ich relacja ze składowymi mechanizmów)
Funkcjonowanie mechanizmów ma odpowiadać za (a dzięki temu wyjaśniać) zjawiska. Mówiąc dość nieprecyzyjnie, przez „zjawisko" stanowiące eksplanandum rozumie się w mechanicyzmie różnego rodzaju własności, funkcje czy zdolności, przypisywane pewnemu złożonemu obiektowi czy systemowi (kategorii obiektów czy systemów). Jest to dość heterogeniczna ontologicznie i nieprecyzyjna charakterystyka. Mówiąc zatem dokładniej, w modelu mechanistycznym przedmiotem wyjaśniania będzie jakiś *wzór lub regularność zachowania* pewnego złożonego obiektu czy systemu. Wedle Roberta Cumminsa (2000) „zjawiska" w istotnym tu sensie należy rozumieć jako „zdolności" (*capacities*) cechujące złożone systemy czy obiekty fizyczne (por.: Glennan 2002; Machamer, Darden, Craver 2011). Zdolności rozumie ten autor z kolei jako własności dyspozycyjne, których naturę można ująć za pomocą przyjmujących postać okresów warunkowych generalizacji, określających jak system będzie się zachowywał przy zajściu danych warunków początkowych. Do „zjawisk" w takim relewantnym dla koncepcji mechanistycznej sensie mogą należeć chociażby mitoza, synteza białek, transmisja synaptyczna, formowanie się gwiazd,

reprodukcja organizmów, transkrypcja i translacja, oksydacja czy pamięć epizodyczna.

Warto na tym etapie wykorzystać pojęcie zjawiska stanowiącego eksplanandum w celu odróżnienia *mechanizmu* tego zjawiska od *systemu* czy *obiektu* fizycznego, który zawiera ten mechanizm. Jeden system czy obiekt – taki, jak choćby żywy organizm lub pojedyncza komórka – może zawierać wiele mechanizmów, z których każdy odpowiada za inną zdolność cechującą ten system czy obiekt. O mechanizmach w istotnym tu sensie nie powinniśmy myśleć jako o odrębnych obiektach, lecz „mechanizmach czegoś" (na przykład mechanizmach mitozy, transmisji synaptycznej czy pamięci epizodycznej). Mechanizmy są zatem częściowo indywiduowane za pomocą zjawisk, za które one odpowiadają i które dzięki temu wyjaśniają (Glennan 2002; Craver 2007: 122–123)[5].

Na czym jednak polega relacja zachodząca między składowymi mechanizmu a zjawiskiem, które jest przez ten mechanizm wyjaśniane? Można rzecz jasna powiedzieć, że mechanizmy „odpowiadają za" albo „stoją u podstaw" zjawisk. Jeszcze inaczej, posiadanie pewnych mechanizmów – pewnej wewnętrznej organizacji strukturalno-funkcjonalnej – przez dany system umożliwia mu posiadanie określonych zdolności. Wielu mechanicystów w taki właśnie lub po-

[5] Należy też zaznaczyć, że nie tylko odkrywanie mechanizmów, ale poprawne charakteryzowanie i odróżnianie podlegających wyjaśnieniu zjawisk jest w praktyce naukowej zasadniczym wzywaniem poznawczym. Odpowiednie wyróżnienie eksplanandum stanowi warunek konieczny powodzenia mechanistycznego wyjaśnienia tego zjawiska. Craver (2007: 123–128) wymienia cały szereg rodzajów błędów związanych z wyróżnianiem zjawisk; błędów polegających, między innymi, na postulowaniu zjawisk nieistniejących (mielibyśmy na przykład z nimi do czynienia, gdybyśmy przyjęli kartezjańską teorię działania układu nerwowego i zdecydowali się odkryć mechanizmy odpowiadające za zjawiska wyróżniane na podstawie tej teorii) czy wprowadzeniu błędnej ich taksonomii (klasyfikowaniu kilku osobnych zjawisk jako jednego albo uznania jednego zjawiska za kilka odrębnych). Dla przykładu – w kognitywistyce błędem byłoby poszukiwanie jednego mechanizmu odpowiadającego za wszelkie zdolności pamięciowe, ponieważ mamy tu do czynienia z szeregiem niezależnych eksplanandów – to znaczy osobnych i niezależnych rodzajów pamięci, takich jak pamięć epizodyczna, semantyczna czy robocza – które wymagają osobnych wyjaśnień, odwołujących się do osobnych mechanizmów (Bechtel 2008: 53–58).

dobny sposób charakteryzuje eksplanacyjnie istotną relację między mechanizmami (komponentami i działaniami składowymi) a wyjaśnianymi przez nie zjawiskami.

Jednakże powyższym sformułowaniom można zarzucić, że są zbyt ogólnikowe i nieprecyzyjne. Na czym bowiem polega „stanie u podstaw" czy też „umożliwianie" określonego zjawiska przez mechanizm? Jest to być może najmniej dopracowana część koncepcji wyjaśniania mechanistycznego, biorąc pod uwagę obecny stan badań. Do tej pory najbardziej filozoficznie dopracowana odpowiedź na to pytanie została zaproponowana przez Cravera (2007). Autor ten charakteryzuje eksplanacyjną relację, o której tu mowa, jako *konstytutywną relewantność*. Relewantność konstytutywna to relacja zachodząca między zachowaniem mechanizmu jako całości (zdolnością, którą wyjaśniamy za pomocą tego mechanizmu) a działaniami jego komponentów. Wyjaśnienie mechanistyczne w ujęciu Cravera odwołuje się właśnie do konstytutywnej relewantności komponentów i ich działań dla zachowania mechanizmu jako całości. Zgodnie z propozycją tego autora działanie realizowane przez komponent jest konstytutywnie relewantne dla zdolności posiadanej przez mechanizm jako całość, jeśli (1) wpływając na zachowanie tego komponentu, wpływamy jednocześnie na zachowanie mechanizmu jako całości (na zjawisko stanowiące eksplanandum), a także (2) wpływając na zachowanie mechanizmu jako całości, wpływamy także na to, jak działa ten komponent[6]. Inaczej mówiąc, konstytutywna relewantność opiera się na tym, że zjawiska oraz składowe odpowiedzialnych

[6] Relewantność konstytutywna nie jest według Cravera relacją przyczynową. Nie ma ona bowiem tych cech, które charakteryzują relacje przyczynowe (Craver 2007: 153–154). Po pierwsze, w przeciwieństwie do przyczynowości, relewantność konstytutywna to relacja symetryczna. Po drugie, zachowanie mechanizmu i działania jego komponentów nie są temporalnie oddzielne, jak dzieje się to w przypadku relacji przyczynowych. Innymi słowy, aktywność mechanizmu jako całości i aktywność jego składowych są jednoczesne, natomiast przyczyna i skutek nie mogą zachodzić jednocześnie. Po trzecie, w przeciwieństwie do przyczynowości, relewantność konstytutywna nie zachodzi między bytami logicznie niezależnymi. Komponenty i ich działania konstytuują bowiem (a nie wywołują przyczynowo) według Cravera wyjaśniane zjawisko.

za nie mechanizmów są *wzajemnie manipulowalne*[7]. Zasadniczą zaletą koncepcji Cravera pozostaje to, że pozwala ona precyzyjnie odróżnić aktywne komponenty mechanizmu od czynników zewnętrznych, które wpływają na mechanizm, lecz nie wchodzą w jego skład. Co prawda zarówno manipulowanie komponentami mechanizmu, jak i czynnikami zewnętrznymi może wpływać na wyjaśniane zjawisko. Jednakże interwencje w działanie mechanizmu jako całości (zjawisko) powinny skutkować tylko zmianami w komponentach mechanizmu (nie powinny modyfikować czynników czy okoliczności zewnętrznych). Warunek *wzajemnej* manipulowalności zostaje spełniony tylko w przypadku rzeczywistych komponentów mechanizmu, ale nie czynników zewnętrznych[8].

Zakończmy na tym etapie omawianie *natury* mechanizmów i zajmijmy się tym, co mechanistyczny model wyjaśniania ma do powiedzenia w kwestiach dotyczących *epistemicznej* praktyki naukowców. Wielu zwolenników mechanistycznej koncepcji wyjaśniania przyjmuje bowiem, że mechanicyzm dość naturalnie łączy się z określonymi tezami dotyczącymi natury praktyki badawczej, a także form,

---

[7] Craver inspiruje się tu „interwencjonistyczną" teorią przyczynowości Jamesa Woodwarda (2003), która zostanie jeszcze przywołana i omówiona w rozdziale 5 (sekcja 5.3.3).

[8] Trzeba tu jednak zaznaczyć, że chociaż propozycja Cravera wydaje się najbardziej dopracowana na tle istniejącej literatury, to z pewnością nie jest ona bezproblemowa. Można jej postawić co najmniej dwa dość mocne zarzuty. Po pierwsze, „wyjaśnianie" czegoś przez coś innego to relacja asymetryczna, a konstytutywna relewantność jest relacją symetryczną (Schindler 2013). Po drugie, niektórzy filozofowie (por.: Leuridan 2012; Schindler 2013) zwracają uwagę na fakt, że stanowisko Cravera implikuje, iż między mechanizmami a ich składowymi zachodzą relacje przyczynowe. Nawet jeśli sama konstytutywna relewantność nie może być jako taka uznana za relację przyczynową (por. przypis 6), to przecież w jej skład wchodzi idea, że możemy (przyczynowo) manipulować komponentami mechanizmu przez manipulowanie mechanizmem jako całością (wyjaśnianym zjawiskiem). To jednak wydaje się niespójne z ideą, że komponenty i działania składowe konstytuują mechanizm, a nie wpływają na niego przyczynowo (Leuridan 2012). Jak zatem widać, oparta na pojęciu wzajemnej manipulowalności koncepcja relewantności konstytutywnej generuje pewne problemy. Na obecnym etapie badań nie istnieje konsensus co do tego, czy koncepcja ta powinna zostać całkowicie odrzucona, czy też można ją obronić, wprowadzając określone poprawki i rewizje.

jakie przyjmują efekty tej praktyki, czyli, mówiąc szeroko, naukowe reprezentacje świata.

Mechanicyzm każe interpretować praktykę badawczą naukowców – a przynajmniej pewną jej część – jako zmierzającą do odkrywania mechanizmów (Craver 2007; Bechtel 2008). Weźmy na przykład pod uwagę opisaną powyżej, opartą na wzajemnej manipulowalności koncepcję relacji zachodzącej między wyjaśnianymi zjawiskami a składowymi mechanizmów. Koncepcja ta pozwala zinterpretować poznawczą rolę, jaką w procesie odkrywania mechanizmów odgrywają badania eksperymentalne. Chodzi tu konkretnie o badania eksperymentalne, które Craver (2007: 144–152) określa eksperymentami międzypoziomowymi. Z jednej strony autor ten wyróżnia eksperymenty „oddolne". Polegają one na tym, że badacz aktywnie wpływa na działanie komponentu mechanizmu – bądź czegoś hipotetycznie postulowanego jako komponent mechanizmu – oraz sprawdza, jak tego rodzaju interwencja wpływa na zdolność czy funkcję przypisywaną mechanizmowi jako całości (na zachodzenie zjawiska stanowiącego eksplanandum)[9]. Z drugiej strony Craver wyróżnia „odgórne" eksperymenty aktywacyjne. W tym przypadku interwencja badacza polega na wpłynięciu nie tyle na składową mechanizmu, co raczej na mechanizm (system zawierający ten mechanizm) jako całość. Eksperymenty odgórne polegają więc na aktywowaniu lub amplifikacji zjawiska stanowiącego eksplanandum i weryfikacji, czy interwencja ta wpływa na zmianę aktywności komponentu – czy też czegoś jedynie postulowanego jako komponent – mechanizmu mają-

[9] Craver (2007: 147–151) wyróżnia dwa typy eksperymentów oddolnych: interferencyjne i stymulacyjne. W przypadku pierwszych badacz zaburza działanie komponentu, natomiast w przypadku drugich – wzmacnia czy pobudza to działanie. W badaniach, które zmierzają do odkrycia mechanizmów stojących u podstaw zjawisk poznawczych, przykładem pierwszej (interferencyjnej) strategii eksperymentalnej są badania oparte na weryfikowaniu wpływu lezji (zastanych lub tymczasowo „spreparowanych" za pomocą przezczaszkowej stymulacji magnetycznej) na funkcjonowanie poznawcze. Przykładem drugiej (stymulacyjnej) strategii mogą być badania oparte na weryfikowaniu wpływu stymulacji elektrycznej określonych fragmentów mózgu na funkcjonowanie poznawcze.

cego wyjaśniać to eksplanandum[10]. Przeprowadzanie obu typów eksperymentów pozwala weryfikować, czy zjawisko wyjaśniane przez mechanizm oraz komponenty i działania tego mechanizmu są wzajemnie manipulowalne, a zatem – czy zachodzi między nimi relacja konstytutywnej relewantności (por. też Baetu 2012). Craver (2007: 144–152) przyznaje, że tak pojęte badania eksperymentalne są jedynie zawodnymi narzędziami, przynoszącymi niejednokrotnie dość kłopotliwe interpretacyjnie wyniki. Problemy te jednak nie są na tyle poważne, by nie można było uznać, że niektóre rodzaje badań eksperymentalnych da się bez trudu zinterpretować „mechanistycznie", to jest jako zabiegi badawcze zmierzające do odkrycia, jakie komponenty i działania składają się na dany mechanizm.

Możliwość aplikowania mechanicyzmu do analizy praktyki badawczej naukowców nie kończy się jednak na rozważaniu roli eksperymentów. Bechtel (2008: 69–71) na przykład zwraca uwagę na rolę *heurystyk* stosowanych przez naukowców w ich próbach odkrycia mechanizmów różnych zjawisk. W praktyce badawczej wykorzystuje się często choćby „heurystyczną teorię identyczności". Korzystanie z tej heurystyki opiera się na przyjmowaniu założenia, że między własnościami znajdującymi się (przynajmniej *prima facie*) na różnych poziomach organizacji lub wyjaśniania zachodzi identyczność (McCauley, Bechtel 2001; Bechtel 2008: 69–71; por. Wimsatt 2006b). Takie „heurystyczne" tezy identycznościowe służą jednak w praktyce nie jako definitywne rozstrzygnięcia jakiegoś problemu, lecz jako użyteczne „narzędzia" ukierunkowujące dalsze badania. Bechtel (2008: 91–105) analizuje rolę heurystycznej teorii identyczności w odkrywaniu mechanizmów odpowiedzialnych za zdolności poznawcze. Pokazuje on pozytywną rolę, jaką w tym procesie

[10] W naukach kognitywnych przykładem tego rodzaju eksperymentów są badania oparte na weryfikacji (na przykład za pośrednictwem funkcjonalnego rezonansu magnetycznego) wpływu realizowanych przez badanego zadań poznawczych na zwiększenie poziomu aktywności określonych obszarów mózgu. Selektywnie aktywowane w ten sposób obszary stają się „kandydatami" do miana funkcjonalnie wyspecjalizowanych komponentów mechanizmu odpowiadającego za tę czy inną zdolność poznawczą (dokładniej: tę zdolność, której wykorzystania wymaga zadanie eksperymentalne realizowane przez badanego).

odegrało przyjmowane często przez badaczy założenie, że określone rodzaje zdolności poznawczych (na przykład percepcja wzrokowa) są „zlokalizowane" w dobrze określonych, wyodrębnionych częściach mózgu (na przykład w obszarze V1 kory wzrokowej). Według Betchtela przyjmowane heurystycznie przez naukowców tezy o zachodzeniu identyczności między rodzajami mentalnymi/psychologicznymi a rodzajami neurobiologicznymi nie mają jednak w nauce statusu analogicznego do podobnych tez identycznościowych stawianych w filozofii umysłu. W praktyce naukowej nie stanowią one nigdy teoretycznego „punktu dojścia", lecz są wstępnymi, „prowizorycznymi", a jednak użytecznymi założeniami, których przyjęcie pozwala na zawężenie i poprawne ukierunkowanie dalszych poszukiwań.

Mechanicyzm niesie też istotne konsekwencje dla filozoficznych rozważań na temat form, jakie przyjmują naukowe *reprezentacje* świata. Koncepcja wyjaśniania mechanistycznego odchodzi od idei, że wiedza naukowa jest zawarta (przede wszystkim) w teoriach przyjmujących postać systemów dedukcyjnych. Naturalnie sprzyja ona za to idei, że duża część wiedzy naukowej jest reprezentowana w postaci *modeli*[11], rozumianych jako takie rodzaje reprezentacji, które (a) są podobne pod pewnymi względami do tego, co reprezentowane (to znaczy odzwierciedlają modelowany obiekt czy proces); (b) mogą być wykorzystywane w celu przeprowadzenia symulacji, w których model podlega manipulacjom, „zastępując" właściwy przedmiot badań. Oczywiście podkreślanie roli, jaką w nauce pełnią tak rozumiane modele, nie jest czymś szczególnie oryginalnym (por.: Giere 2004, 2010; Grobler 2006: 175–178). Jednakże mechanicyzm pozwala dookreślić funkcję modeli oraz rozjaśnić, dlaczego są one tak ważne i rozpowszechnione w nauce. Jak bowiem zauważa Peter Godfrey-Smith (2005, 2006), budowanie modeli to naturalny i wygodny sposób reprezentowania *mechanizmów*. Modele dobrze nadają się do tego, by odzwierciedlać złożone, fizyczne struktury występujące w naturze. Praktyka eksplanacyjna naukow-

[11] Mowa tu rzecz jasna o modelach jako tworzonych wolicjonalnie i świadomie, publicznie dostępnych, naukowych reprezentacjach świata, a nie o wewnętrznych, subosobowych modelach, o których będzie mowa w rozdziale 4.

ców polega według tego autora często na tworzeniu hipotetycznych struktur mających docelowo odzwierciedlać swoją organizacją (modelować) rzeczywistą organizację mechanizmów stojących u podstaw poszczególnych zjawisk. Każdy taki model będzie początkowo niekompletny i pełen idealizacji, jednak będzie on stopniowo rewidowany i uzupełniany tak, by ostatecznie odzwierciedlał on rzeczywistą strukturę i organizację modelowanego mechanizmu. Craver (2007: 113–114) ów pierwotny, niekompletny model mechanizmu nazywa „szkicem" (*mechanism sketch*), natomiast model uzupełniony i kompletny (aspirujący do miana kompletnego) – „schematem" mechanizmu (*mechanism schema*). Proces budowania wyjaśnień mechanistycznych opiera się na stopniowym przechodzeniu od szkiców mechanizmów do ich schematów.

Mechanistyczny model wyjaśniania pozwala też zwrócić uwagę na pomijaną często w filozofii nauki kwestię roli pełnionej przez *ikoniczne*, a nie jedynie językowe reprezentacje świata. Według Bechtela (2008: 17–22) często spotykane w artykułach naukowych diagramy stanowią zazwyczaj reprezentacje postulowanych przez badaczy mechanizmów. Połączone strzałkami elementy diagramu mają niejednokrtonie odzwierciedlać organizację komponentów i operacji składowych. Powszechna obecność tego rodzaju reprezentacji w nauce jest spowodowana właśnie tym, że stanowią one przejrzysty i łatwy do zrozumienia sposób reprezentowania mechanizmów. Co istotne, zdaniem Bechtela tego rodzaju ikoniczne reprezentacje nie są jedynie narzędziem pomocniczym, dodatkiem do reprezentacji przyjmujących formę lingwistyczną, lecz stanowią niezależny i pełnoprawny sposób reprezentowania świata przez naukowców. Nie znaczy to rzecz jasna, że naukowe reprezentacje mechanizmów nie mogą przyjmować formy opisu zakodowanego w języku naturalnym. Funkcjonowanie mechanizmów można jak najbardziej opisać językowo. Chodzi raczej o to, że status takich opisów w koncepcji wyjaśniania mechanistycznego nie jest zasadniczo różny od statusu, który przysługuje reprezentacjom ikonicznym[12].

[12] Jaka jest relacja między modelami mechanizmów a ikonicznymi oraz nieikonicznymi (językowymi) reprezentacjami mechanizmów w nauce? Proponuję

2.1.2. Mechanicyzm, nomologiczno-dedukcyjny model wyjaśniania i pluralizm eksplanacyjny

Jak mechanicyzm ma się do alternatywnych, rozwijanych przez filozofów nauki koncepcji wyjaśniania naukowego? Pozostając na wysokim poziomie ogólności, można przyjąć *erotetyczne* kryterium odróżniające koncepcję mechanistyczną od pozostałych teorii wyjaśniania (a w każdym razie wielu spośród nich). Otóż niektóre filozoficznie wpływowe sposoby myślenia o naturze wyjaśniania w nauce są inspirowane ideą, że wyjaśnienie określonego zjawiska *Z* polega na udzieleniu odpowiedzi na pytanie: „Dlaczego zaszło (zachodzi, zajdzie) *Z*?" (por. Grobler 2006: 112). Poszczególne teorie wyjaśniania mówią co innego o tym, jak powinna wyglądać odpowiedź na tak postawione pytanie. Klasyczna, nomologiczno-dedukcyjna (omówiona szerzej poniżej) teoria wyjaśniania postuluje, że odpowiedź powinna odwoływać się do praw naukowych oraz poprzedzających zjawisko *Z* warunków początkowych. Zgodnie z tradycją wywodzącą się z prac Wesleya Salmona (1971, 1984) odpowiedź na pytanie: „Dlaczego zaszło (zachodzi, zajdzie) *Z*?", powinna odwoływać się raczej do statystycznie istotnych czynników prowadzących do zajścia *Z* lub do przyczyn zajścia *Z*. Jaki rodzaj odpowiedzi proponuje zatem koncepcja wyjaśniania za pomocą mechanizmów?

Proponuję przyjąć, iż mechanicyzm wyróżnia się na tle wymienionych koncepcji wyjaśniania nie tyle tym, że postuluje alternatywny rodzaj odpowiedzi na pytanie „dlaczego?", ile tym, że ujmuje wyjaśnianie jako odpowiedź na inny rodzaj pytania. Mianowicie, jak już wcześniej zaznaczyłem, wyjaśnienia mechanistyczne mają stanowić odpowiedzi na pytania „jak?". Dokładniej, wyjaśnienie mechani-

przyjąć, że modele są kategorią nadrzędną w stosunku do diagramów oraz innych (w tym nieikonicznych) form, jakie mogą przyjmować naukowe reprezentacje mechanizmów. Precyzyjniej, modele naukowe to abstrakcyjne, często wyidealizowane reprezentacje organizacji mechanizmów, natomiast diagramy i inne reprezentacje ikoniczne (jak również nieikoniczne) to sposoby wyrażania lub kodowania modeli w pewnej konkretnej, dostępnej percepcyjnie postaci (por. Godfrey-Smith 2006). Z takiego punktu widzenia praktyka poszukiwania mechanizmów jest ściśle powiązana z praktyką budowania modeli mechanizmów, a modele te często są wyrażone w postaci ikonicznej.

styczne powinniśmy rozumieć jako odpowiedź na pytanie: „Jak system *S* wykonuje *Z*?", gdzie zjawisko *Z* to zdolność posiadana przez fizyczny system *S*. Zgodnie z mechanicyzmem odpowiedź na tego rodzaju pytanie będzie przyjmowała postać reprezentacji (modelu) mechanizmu, którego działanie jest odpowiedzialne za to, że *S* ma stanowiącą eksplanandum zdolność *Z*.

Erotetyczne kryterium pozostaje rzecz jasna bardzo „gruboziarniste". Nadal można zapytać, na czym polegają inne, bardziej szczegółowe różnice między mechanicyzmem a alternatywnymi modelami wyjaśniania. To niezwykle szeroki temat. Zamiast podejmować go w sposób wyczerpujący, pragnę skupić się jedynie na różnicach zachodzącym między koncepcją mechanistyczną a najbardziej klasycznym, nomologiczno-dedukcyjnym (N-D) modelem wyjaśnienia naukowego (Hempel, Oppenheim 1948; Nagel 1970). Dlaczego akurat nim? Wybór ten jest podyktowany faktem, że to właśnie model N-D najbardziej wpłynął na sposób, w jaki filozofowie umysłu i kognitywistyki traktowali (traktują) cały szereg problemów bezpośrednio związanych z przedmiotem tej książki, czyli z zagadnieniem reprezentacji mentalnych. Model N-D naturalnie wiąże się bowiem z określonym postrzeganiem między innymi relacji międzypoziomowych w nauce, natury wyjaśniania redukcyjnego i eliminacji czy też z ideą psychologii potocznej jako teorii wyjaśniającej ludzkie działania za pomocą praw. Inspirowane modelem N-D podejście do tych zagadnień motywowało z kolei koncepcje dotyczące natury reprezentacji mentalnych i ich statusu eksplanacyjnego – w szczególności statusu eksplanacyjnego postaw propozycjonalnych – które będę w tej pracy (konkretnie w rozdziale 5) krytykował i próbował zastąpić propozycjami alternatywnymi, inspirowanymi modelem mechanistycznym. Żadna inna teoria wyjaśniania naukowego – jak chociażby wspomniany model Salmona (1971, 1984) czy unifikacyjny model Philipa Kitchera (1989) – nie miała nawet po części tak znacznego wpływu na myślenie o reprezentacjach, jak model N-D. Dlatego też odniesienie się właśnie do tego ostatniego jest strategicznie istotne dla celów tej pracy.

Te różnice zachodzące między mechanicyzmem a modelem N-D, które dotyczą postrzegania redukcji oraz relacji międzypozio-

mowych, zostaną szerzej omówione w następnym podrozdziale. Na obecnym etapie rozważań warto zaznaczyć dwa inne ważne punkty, w których mechanicyzm i model N-D są zasadniczo odmienne.

Po pierwsze, obie koncepcje wyjaśniania naukowego różnią się, jeśli chodzi o postrzeganie relacji między wyjaśnianiem a *przewidywaniem*. Zasadniczą ideą leżącą u podstaw modelu N-D jest teza, iż wyjaśnianie zjawiska polega na wskazaniu, że zdanie opisujące zajście tego zjawiska wynika dedukcyjnie z koniunkcji złożonej ze zdania specyfikującego stosowne prawo naukowe oraz zdania specyfikującego warunki początkowe. Co charakterystyczne, zgodnie z tym modelem można analogicznie opisać to, w jaki sposób naukowcy przewidują zachodzenie określonych zjawisk. Podobnie jak wyjaśnianie, predykcja polega na dedukcji opartej o parę zdań opisujących stosowne prawo naukowe oraz warunki początkowe. Tym samym wyjaśnienie zjawiska w modelu N-D to pokazanie, że jego zajście może być racjonalnie oczekiwane czy przewidywane. Mechanicyzm traktuje z kolei przewidywanie i wyjaśnianie jako oddzielne i względnie niezależne (por.: Cummins 2000; Godfrey-Smith 2005; Craver 2007: 39–40). Dysponowanie mechanistycznym wyjaśnieniem zachowania pewnego systemu nie jest ani wystarczające, ani konieczne do tego, byśmy byli zdolni to zachowanie przewidywać. Jak zauważa Cummins (2000), często znamy i rozumiemy mechanizm czy mechanizmy odpowiadające za zachowanie systemów na tyle złożonych, że ich działanie w praktyce nie poddaje się predykcji. Jednocześnie – często potrafimy przewidywać zachowanie różnych systemów w przyrodzie nawet wtedy, gdy nie znamy mechanizmu odpowiedzialnego za to zachowanie. Na przykład ludzie byli zdolni przewidywać pływy mórz, jeszcze zanim zrozumieli odpowiadający za te pływy mechanizm (Cummins 2000). Ściśle ze sobą splecione w modelu N-D kategorie predykcji i wyjaśniania są w koncepcji mechanistycznej wyraźnie odrębne[13].

[13] Peter Godfrey-Smith (2005) ostrzega jednak przed przesadnym podkreślaniem tej odrębności. Choć wyjaśnianie mechanistyczne nie wymaga przewidywania jako warunku koniecznego, to w praktyce powinniśmy oczekiwać, że poprawne wyjaśnienia pozwolą nam na sformułowanie pewnych przewidywań dotyczących zachowania badanego systemu (w tym jego zachowania w nowych

Po drugie, mechanistyczna koncepcja wyjaśniania wyraźnie odbiega od modelu N-D w kwestii eksplanacyjnej roli *praw naukowych*. Z punktu widzenia modelu N-D prawa naukowe odgrywają fundamentalną rolę w wyjaśnianiu. W mechanicyzmie to twierdzenie jest odrzucane (Cummins 2000; Craver 2007: 66–69; Bechtel 2008: 142––143). Zwolennicy modelu mechanistycznego zwracają często uwagę na fakt, że w biologii czy naukach inżynieryjnych wyjaśnienia odwołujące się do praw są rzadkie i nie mają centralnego znaczenia. Powodem takiego stanu rzeczy ma być fakt, że w naukach tych rzeczywista praktyka eksplanacyjna po prostu odbiega od tego, jak ją opisuje model N-D. Zgodnie z koncepcją mechanistyczną wyjaśnienia w tych dyscyplinach przyjmują w większości postać wyjaśnień mechanistycznych, a te nie odwołują się do praw naukowych. Raz jeszcze warto powołać się na obserwacje poczynione przez Cumminsa (2000). Autor ten zwraca uwagę na fakt, że jeśli badacze w naukach stosujących wyjaśnienia mechanistyczne w ogóle powołują się na prawa (lub prawopodobne generalizacje), to te nie występują w roli eksplanansów, lecz *eksplanandów*. Odwołując się do nauk kognitywnych, Cummins podaje przykład efektu McGurka, czyli zjawiska związanego z interakcjami różnych modalności zmysłowych[14]. Otóż efekt ten może zostać scharakteryzowany jako określonego rodzaju prawo specyfikujące zachodzenie pewnych ogólnych, systematycznych zależności między dwiema modalnościami zmysłowymi. W psychologii powszechnie przyjmuje się jednak, że ma ono status eksplanandum, a nie eksplanansu. Efekt McGurka w praktyce eksplanacyjnej naukowców jest traktowany nie jako podstawa do wyjaśniania jakiegokolwiek zjawiska, lecz raczej opis zjawiska, które domaga się wyjaśnienia w kategoriach mechanistycznych. Uogólniając,

warunkach, w których na ogół nie znajduje się on w naturze). Nie jest więc tak, że przewidywanie i wyjaśnianie są w mechanicyzmie kompletnie niepowiązane; chodzi jedynie o to, że nie są one ze sobą związane tak ściśle, jak w modelu N-D: wyjaśnianie nie implikuje przewidywania i *vice versa*.

[14] Efekt ten polega na tym, że percepcja słuchowa sylab przez daną osobę podlega systematycznym zmianom, jeśli osoba słuchająca jednocześnie percypuje wzrokowo drugą osobę, wypowiadającą sylabę inną niż słyszana (choć fonetycznie do niej zbliżoną).

prawa nadają się do *opisania* zdolności przysługujących systemowi poznawczemu (czy jakiemukolwiek systemowi, którego zachowanie jest wyjaśniane mechanistycznie), ale już mechanistyczne *wyjaśnienia* tych zdolności odwołują się nie do praw naukowych, lecz do zorganizowanych, działających komponentów mechanizmów. W wyjaśnieniach mechanistycznych prawa (prawopodobne generalizacje) nie są eksplanansami, lecz dostarczają opisów eksplanandów[15].

Zanim można będzie przejść dalej, należy poczynić jeszcze jedną uwagę. Choć mechanistyczny model wyjaśniania odgrywa w tej książce zasadniczą rolą, moją intencją nie jest akceptacja mocnej tezy, że wszelkie wyjaśniania naukowe mają charakter wyjaśnień mechanistycznych. Niewykluczone, że żaden pojedynczy model wyjaśniania sformułowany przez filozofów nauki – w tym model mechanistyczny – nie opisuje bez wyjątku całości praktyki eksplanacyjnej naukowców. Wydaje się, że pozycja pluralistyczna w tej sprawie jest bardziej wiarygodna. Zgodnie z nią pełnoprawnie naukowe wyjaśnienia mogą przyjmować różną postać, w zależności od dziedzi-

[15] Nawiasem mówiąc, jeden z anonimowych recenzentów tej książki zasugerował, że zachodzi jeszcze jedna ważna różnica między mechanicyzmem a modelem N-D. Poprawność wyjaśnienia mechanistycznego zależy nie tylko od jego wartości predykcyjnej, ale też od tego, czy poprawnie opisuje ono strukturalną i funkcjonalną organizację pewnego realnego mechanizmu. Fakt, iż poprawność wyjaśnień mechanistycznych zależy od struktury świata, miałby właśnie stanowić kolejny czynnik odróżniający wyjaśnienia mechanistyczne od wyjaśnień N-D – w tych ostatnich powodzenie wyjaśniania zależy bowiem tylko od zależności logicznej między przesłankami a wnioskiem. Innymi słowy, wyjaśnienia mechanistyczne mają wbudowane założenia czy zobowiązania metafizyczne, które nie towarzyszą wyjaśnieniom odwołującym się do praw. Wartość wyjaśnień mechanistycznych nie zależy tylko od ich „zalet" czysto epistemologicznych czy logicznych, ale też od budowy realnych mechanizmów. To bardzo ciekawa uwaga, jednak mam w stosunku do niej pewną wątpliwość. W filozofii nauki istnieją realistyczne interpretacje praw w nauce, zgodnie z którymi (przynajmniej niektóre) prawa naukowe nie tylko pozwalają na przewidywanie zjawisk, ale także opisują prawa w jakimś sensie rzeczywiście „rządzące" przyrodą (por. Carroll 2010). N-D model wyjaśniania naukowego można zatem uzupełnić twierdzeniem, że warunkiem poprawności wyjaśnień za pomocą praw jest nie tylko poprawność logiczna, ale też realność – jakkolwiek ją rozumieć – stosownego prawa (czy może obiektów/własności/procesów denotowanych przez terminy czy predykaty, za pomocą których to prawo jest sformułowane).

ny naukowej lub wyjaśnianego zjawiska. Na przykład jeśli potraktujemy fizykę jako naukę fundamentalną, w której częstokroć nie jest możliwe odkrywanie mechanizmów znajdujących się na niższym poziomie organizacji niż eksplanandum – ponieważ taki poziom nie istnieje lub nic o jego istnieniu na danym etapie rozwoju wiedzy naukowej nie wiemy – to trudno uznać, że formułowane w jej ramach wyjaśnienia są mechanistyczne. Do opisu praktyki eksplanacyjnej fizyków może się zatem dobrze nadawać stawiający w swoim centrum prawa naukowe model N-D. Zarazem znaczna część wyjaśnień formułowanych w biologii albo kognitywistyce może mieć mechanistyczny charakter[16].

Co więcej, nawet jeśli zrelatywizujemy swoje rozważania do jednej dyscypliny, w obrębie której wyjaśniamy jedno i to samo eksplanandum, wcale nie musimy uznać, że zjawisko to może być wyjaśnione jedynie mechanistycznie. Dana dyscyplina może chociażby wyjaśniać dane zjawisko zarówno przez opisanie jego mechanizmu, jak i za pomocą (dajmy na to) etiologicznego wyjaśnienia przyczynowego, czyli przez wskazanie przyczyny tego zjawiska, która (1) czasowo poprzedza to zjawisko; (2) występuje na tym samym poziomie organizacji – a nie na poziomie niższym, jak komponenty i działania składowe mechanizmów – co to zjawisko (por. Craver 2007: 74, 93––104[17]). Na przykład biologów może interesować nie tylko mechanizm odpowiadający za fotosyntezę, ale także środowiskowe czyn-

[16] Można oczekiwać, że praktyka wyjaśniania za pomocą mechanizmów będzie szczególnie ważna we wszystkich dyscyplinach, które mają inżynieryjny charakter. Mowa tu zarówno o dyscyplinach zajmujących się projektowaniem złożonych systemów, jak i dyscyplinach dokonujących „inżynierii odwrotnej" systemów już istniejących – czy to naturalnych, czy artefaktualnych.

[17] Taka koncepcja etiologicznego wyjaśniania przyczynowego opiera się na pracach Salmona (w szczególności: 1984). Należy podkreślić, że ten typ wyjaśniania różni się od wersji mechanistycznej tym, że skupia się on na interakcjach przyczynowych w obrębie jednego poziomu organizacji. W wyjaśnianiu mechanistycznym istotne jest z kolei nie tylko opisanie zależności przyczynowych między komponentami mechanizmu, ale też wskazanie, jak zorganizowane (również przyczynowo) operacje tych komponentów wywołują zjawisko wyróżniane na wyższym poziomie organizacji. Co ciekawe, choć sam Salmon (1984) posługiwał się terminem „mechanizm przyczynowy", to nie zdefiniował go on w sposób, który by odpowiadał temu, jak mechanizmy są rozumiane

niki zewnętrzne wpływające przyczynowo na zainicjowanie tego procesu. Pełne wyjaśnienie zjawiska biologicznego może zatem wymagać zarówno wskazania odpowiadającego za nie mechanizmu, jak i jego przyczynowej etiologii. Jeden typ wyjaśnienia odpowiada na pytanie o „jak" zjawiska, a drugi – na pytanie o to, „dlaczego" ono zachodzi. Są to odpowiedzi komplementarne, a nie alternatywne.

Zgodnie zatem z duchem pluralizmu eksplanacyjnego, mechanicyzm będę tu uznawał za jedną z potencjalnie wielu form wyjaśniania naukowego. Przyjmuję jednak zarazem, że w konkretnej nauce stanowiącej przedmiot zainteresowania tej książki – w kognitywistyce – zasadniczą czy dominującą (nawet jeśli nie jedyną) formą wyjaśniania jest wyjaśnianie przez wskazywanie mechanizmów zjawisk. Kognitywistyka przypomina pod tym względem bardziej biologię niż fizykę. Do myśli tej powrócę w sekcji 2.3.

2.2. Wyjaśnianie mechanistyczne a relacje międzypoziomowe

2.2.1. Poziomy w nauce jako poziomy mechanizmów

Przyjrzyjmy się teraz bliżej zaznaczonej już powyżej kwestii wielopoziomowej natury mechanizmów (oraz wyjaśnień mechanistycznych). Mechanistyczne pojmowanie relacji międzypoziomowych w nauce jest szczególnie istotne w kontekście tej pracy, ponieważ zostanie ono wykorzystane w rozwijanej w rozdziale 5 argumentacji na rzecz mechanistycznej neutralności psychologii potocznej. Precyzyjniej, będę tam zmierzać do pokazania, że aparatura pojęciowa psychologii potocznej ma zastosowanie na najwyższym poziomie mechanistycznej organizacji systemu poznawczego, mianowicie na poziomie tego systemu jako całości zaangażowanej w interakcje ze swoim środowiskiem. Teza ta zaś będzie się odwoływała do założenia, że relacje między poziomami organizacji systemu poznawczego (oraz skorelowanymi z nimi sposobami opisywania i wyjaśniania

we współczesnych koncepcjach wyjaśniania mechanistycznego (por.: Glennan 2002: S343).

działania tego systemu) powinny być rozumiane właśnie jako relacje między poziomami w hierarchicznej organizacji mechanizmów.

„Poziomy" w koncepcji mechanistycznej to poziomy hierarchicznej („szkatułkowej") organizacji mechanizmów. Są one powiązane określonego rodzaju relacją kompozycji (część–całość). Komponenty mechanizmu i wykonywane przez nie działania znajdują się na niższym poziomie niż mechanizm jako całość (czy też system fizyczny zwierający ten mechanizm). Bardziej technicznie:

> *X* wykonujący operację φ znajduje się na niższym poziomie, niż *S* wykonujący operację ψ wtedy i tylko wtedy, gdy *x* wykonujący operację φ stanowi komponent w mechanizmie odpowiadającym za wykonywanie ψ przez *S* (za: Craver 2007: 189).

Taką pozytywną charakterystykę poziomów w wyjaśnianiu mechanistycznym warto uzupełnić charakterystyką o charakterze negatywnym, czyli taką, która *explicite* pokazuje, czym poziomy mechanizmów *nie* są. W nauce i filozofii nauki poziomy – rozumiane jako poziomy czy to organizacji samej nauki, czy też badanej przez nią rzeczywistości – są pojmowane na wiele różnych sposobów. Mechanistyczne ujęcie poziomów odbiega od tych bardziej lub mniej rozpowszechnionych sposobów ich rozumienia. Zestawienie mechanistycznej koncepcji poziomów z ujęciami alternatywnymi może pozwolić na dookreślenie tej pierwszej oraz zaznaczenie jej cech specyficznych. Poniższa lista stanowi właśnie tego rodzaju zestawienie. Przedstawia ona wyróżnione przez Cravera (2007: 170–194), obecne (bardziej lub mniej *explicite*) w nauce koncepcje czy sposoby rozumienia poziomów, które pod istotnymi względami różnią się od tego, jak poziomy są rozumiane w mechanicyzmie.

a) Poziomy nauki

- **Natura poziomów w tym znaczeniu**

Są to poziomy organizacji badań naukowych (na przykład poszczególne dyscypliny, programy badawcze, paradygmaty) lub związanych z nimi produktów epistemicznych nauki (na przykład teorii, modeli, wyjaśnień). Poszczególne poziomy organizacji badań lub ich

epistemiczne produkty mają odpowiadać kolejnym „warstwom" organizacji rzeczywistości fizycznej. Nauka stanowiłaby w takiej perspektywie „hierarchiczną" organizację dyscyplin (od fizyki po socjologię i nauki o kulturze) formułujących teorie (modele, wyjaśnienia i tak dalej) mające za swój przedmiot kolejno: cząstki elementarne, atomy, cząsteczki, komórki organizmów, organizmy jako całości oraz systemy społeczno-kulturowe (por. Oppenheim, Putnam 1958).

- **Różnice w stosunku do poziomów mechanistycznych**

(1) W mechanicyzmie odrzuca się założenie, że zachodzi korespondencja między poziomami organizacji lub produktami epistemicznymi nauki z jednej strony, a poziomami organizacji świata fizycznego jako całości z drugiej. Na przykład teorie z zakresu biologii mogą dotyczyć całego szeregu „poziomów" rzeczywistości fizycznej, od poziomu molekularnego po poziom całych ekosystemów. (2) Poziomy w wyjaśnianiu mechanistycznym są zrelatywizowane do poszczególnych mechanizmów czy kategorii mechanizmów. Wyróżnianie poziomów mechanistycznych ma sens tylko w kontekście mechanistycznej (strukturalno-funkcjonalnej) dekompozycji danego systemu, więc poziomy mechanizmów nie są „warstwami" rzeczywistości fizycznej jako takiej.

b) Poziomy przyczynowe (przetwarzania i kontroli)

- **Natura poziomów w tym znaczeniu**

Poziomy przetwarzania to kolejne stadia przetwarzania informacji (na przykład sekwencja czasowa, w której informacje wzrokowe są przetwarzane w kolejnych partiach mózgu). Poziomy kontroli to poziomy podporządkowania, w których wyższe poziomy kontrolują lub regulują działania poziomów niższych (na przykład kontrolowanie określonych partii mózgu przez ośrodki przedczołowe).

- **Różnice w stosunku do poziomów mechanistycznych**

(1) Między poziomami przyczynowymi (przetwarzania i kontroli) nie zachodzi relacja kompozycyjna, natomiast poziomy mechanizmów stanowią formę poziomów kompozycyjnych. (2) W przeciwieństwie do poziomów przyczynowych – procesy zachodzące na poziomach mechanizmów nie zachodzą sekwencyjnie (działanie mechanizmu jako całości oraz działanie jego komponentów jest jednoczesne).

(3) Relacja między poziomami mechanizmów nie jest relacją przyczynową, podczas gdy poziomy przetwarzania i kontroli wchodzą w relacje przyczynowe.

c) Poziomy realizacji

- **Natura poziomów w tym znaczeniu**

Własności wyższego poziomu to własności funkcjonalne. Własności niższego poziomu to własności fizyczne, które realizują własności funkcjonalne wyższego poziomu. Na przykład własności neurobiologiczne (niższy poziom) jako realizujące funkcjonalnie rozumiane własności mentalne (wyższy poziom)[18].

- **Różnice w stosunku do poziomów mechanistycznych**

Poziomy realizacji nie są powiązane kompozycyjnie, natomiast relacja między poziomami mechanizmów to relacja kompozycji. (Por. też omówienie zagadnienia wielorakiej realizacji w kontekście mechanicyzmu zawarte w sekcji 2.3.2).

d) Poziomy rozmiaru

- **Natura poziomów w tym znaczeniu**

Poziomy wielkości fizycznej obiektów. Obiekty mniejsze byłyby zlokalizowane na niższym poziomie niż obiekty większe, przy czym w ramach poszczególnych poziomów wielkości dochodzi do maksymalnej regularności i przewidywalności interakcji przyczynowych (Wimsatt 1976; za: Craver 2007: 180–182)[19]. Na przykład (1) cały Wszechświat może być rozumiany jako „warstwowo" ułożony na różnych poziomach wielkości: od poziomu subatomowego po po-

[18] Poziomami realizacji są chociażby trzy poziomy wyjaśniania wyróżnione przez Davida Marra (2010: 24– 27): obliczeniowy, algorytmiczny/reprezentacyjny i implementacyjny. Na poziomie obliczeniowym jest określana funkcja wykonywana przez system. Na poziomie algorytmicznym/reprezentacyjnym są określane algorytmy oraz reprezentacje wykorzystywane w celu obliczenia tej funkcji – innymi słowy, realizujące tę funkcję. Na poziomie implementacyjnym są określane struktury neuronalne realizujące (implementujące) algorytmy wyróżnione na poziomie algorytmicznym/reprezentacyjnym.

[19] Obiekty w obrębie jednego poziomu wchodzą w interakcje przyczynowe częściej i w bardziej regularny sposób, niż jest to w przypadku interakcji obiektów występujących na różnych poziomach wielkości.

ziom obiektów kosmologicznych; (2) ośrodkowy układ nerwowy może być postrzegany jako złożony na różnych poziomach wielkości: od poziomu molekularnego, przez poziom pojedynczych neuronów, poziom systemów neuronalnych, aż po ośrodkowy układ nerwowy jako całość.

- **Różnice w stosunku do poziomów mechanistycznych**

(1) Interakcje zachodzące między obiektami znacznie różniącymi się wielkością (na przykład między cząsteczkami a całymi komórkami, między wirusami a całymi organizmami) mogą cechować się wysokim stopniem regularności i przewidywalności interakcji. Dlatego też obiekty znacznie różniące się wielkością mogą stanowić komponenty jednego mechanizmu, to jest znajdować się na jednym poziomie mechanistycznym (na przykład wirusy oraz całe populacje organizmów w ekosystemie). (2) Między poziomami organizacji mechanizmów zachodzą co prawda różnice wielkości (mechanizmy jako całości są większe od ich komponentów), jednak są one jedynie pochodną konsekwencją kompozycyjnej natury mechanizmów.

e) Poziomy mereologiczne

- **Natura poziomów w tym znaczeniu**

Kompozycyjne poziomy w ramach relacji część–całość, definiowanej przez mereologię jako naukę formalną.

- **Różnice w stosunku do poziomów mechanistycznych**

Relacja część–całość w mechanizmach nie spełnia szeregu twierdzeń zawartych w większości systemów mereologicznych. Relacja między mechanizmami a ich komponentami nie jest na przykład samozwrotna (mechanizmy nie są własnymi częściami) ani ekstensjonalna (zjawiska z wyższych poziomów organizacji nie stanowią funkcji samych elementów składowych mechanizmu – w mechanizmach zasadniczą rolę odgrywa *organizacja* komponentów i działań).

f) Poziomy agregacji

- **Natura poziomów w tym znaczeniu**

Poziomy są wyznaczane przez relację część–całość (kompozycyjną), gdzie całość (własność całości) stanowi agregat, to jest sumę swoich

części (własności części). Na przykład masa obiektu stanowi sumę masy jego części.

- **Różnice w stosunku do poziomów mechanistycznych**

Mechanizmy nie są agregatami. Ich części składowe – komponenty i ich operacje – charakteryzują się nieagregatywną organizacją, w związku z czym w mechanizmach własności całości nie są sumami własności egzemplifikowanych przez części (por. omówienie nieagregatywności mechanizmów w następnej sekcji).

g) Poziomy zawierania przestrzennego

- **Natura poziomów w tym znaczeniu**

Poziomy kompozycyjne w ramach relacji część–całość, gdzie części zawierają się w granicach przestrzennych całości. Jednostkę na niższym poziomie niż całość (cały obiekt czy system) stanowi dowolny przestrzenny „wycinek" tej całości.

- **Różnice w stosunku do poziomów mechanistycznych**

Zawieranie się przestrzennie w mechanizmie nie wystarcza do bycia jego komponentem. Komponenty mechanizmów to nie dowolne elementy zawierające się przestrzennie w mechanizmie, lecz tylko te elementy, które wykonują określone operacje (pełnią funkcje) w ramach tego mechanizmu.

Jak widać w powyższym zestawieniu, mechanistyczne postrzeganie poziomów jest specyficzne i zasadniczo różni się od wymienionych alternatyw[20]. Warto szczególnie podkreślić dwa fakty dotyczące rozumienia poziomów w mechanicyzmie. Po pierwsze, mówiąc o poziomach kompozycyjnej organizacji mechanizmów, nie ma się na

[20] Warto jednakże zaznaczyć, że mechanistyczna wizja relacji międzypoziomowych wykazuje wyraźną zbieżność z wizją zawartą w ogólnej teorii systemów (por. omówienie tej ostatniej w: Poczobut 2009: 344–372). Na przykład – podobnie jak mechanicyzm – teoria systemów kładzie nacisk na znaczenie nieagregatywnej organizacji złożonych struktur występujących w przyrodzie; podobnie jak w mechanicyzmie (por. sekcja 2.2.2), w ramach ogólnej teorii systemów rozpoznaje się też znaczenie faktu, że działanie systemu jako całości nie jest zdeterminowane wyłącznie/całkowicie przez zachowanie jego komponentów (podsystemów). Dziękuję Robertowi Poczobutowi za zwrócenie mi uwagi na tę zbieżność.

myśli poziomów organizacji natury czy też świata fizycznego jako całości. Dowolny komponent mechanizmu znajduje się na niższym poziomie niż mechanizm jako całość tylko relatywnie do określonej mechanistycznej dekompozycji danego systemu. Poza kontekstem konkretnego wyjaśnienia mechanistycznego stratyfikowanie wedle poziomów mechanizmów traci sens. Nie można powiedzieć, że serce znajduje się na „niższym" lub „wyższym" poziomie niż Droga Mleczna albo procesor w komputerze osobistym, ponieważ nie stanowią one komponentów w jednym mechanizmie (Craver 2007: 188–194). Po drugie, poziomy mechanizmów należą do ogólnej kategorii poziomów kompozycyjnych. Są to zatem przede wszystkim poziomy rozumiane metafizycznie (jako poziomy kompozycyjnej organizacji), a nie epistemicznie (jako poziomy opisu i wyjaśniania). Co więcej, sposób rozumienia relacji część–całość w mechanicyzmie jest specyficzny i nie da się go sprowadzić do innych, wskazanych wyżej sposobów. Specyfika komponentów mechanizmów polega na tym, że występują one zawsze w ramach organizacji pewnego mechanizmu i wykonują w mechanizmie określone operacje. Dlatego na przykład nie każdy sposób wyodrębniania części na podstawie ich zawierania się przestrzennego w pewnej całości pozwoli poprawnie wyodrębnić akurat te części, które stanowią aktywne komponenty mechanizmu odpowiadającego za dane zjawisko.

2.2.2. Relacje międzypoziomowe a granice mechanistycznej redukcji

Jedną z ważnych i wpływowych w filozofii nauki idei sformułowanych na gruncie nomologiczno-dedukcyjnego modelu wyjaśniania stanowi twierdzenie, że poziomy w nauce są, lub powinny być, wzajemnie redukowalne, gdzie redukcję rozumie się jako redukcję *interteoretyczną* (por. Nagel 1970). W najbardziej klasycznym ujęciu redukcja tego rodzaju to procedura polegająca na (1) sformułowaniu praw stanowiących logiczne (definicyjne) „mosty" między poszczególnymi predykatami teorii redukowanej i redukującej; (2) pokazaniu, że jest możliwe logiczne (dedukcyjne) wywiedzenie praw teorii redukowanej z praw teorii redukującej. W idealnej sytuacji efektem takiej procedury powinna być możliwość wyjaśnienia za pomocą

teorii redukującej wszystkich zjawisk, które były wcześniej wyjaśniane przez teorię redukowaną. Biorąc pod uwagę, że redukcja interteoretyczna zachodzi między epistemicznymi produktami nauki (teoriami), stanowi ona formę redukcji *reprezentacyjnej*[21].

Jak się wydaje, wyjaśnianie mechanistyczne można również uznać za *pewną* formę wyjaśniania redukcyjnego (por.: Godfrey-Smith 2005; Bechtel 2008: 142–143). Wyjaśnienia mechanistyczne pokazują, w jaki sposób zdolności wyróżniane na poziomie systemowym są umożliwiane przez komponenty i operacje z niższego poziomu. Redukcyjność tych wyjaśnień polega właśnie na wyjaśnianiu wyższych poziomów za pomocą niższych. Zachowania całości są „sprowadzane" do działań zorganizowanych komponentów mechanizmu. Takie postrzeganie redukcji różni się jednak od opisanej wyżej, wywodzącej się z modelu N-D koncepcji redukcji interteoretycznej. Różnice dotyczą dwóch zasadniczych aspektów: (1) tego, co wchodzi w relację bycia redukowanym (natury argumentów relacji redukcji); (2) tego, na czym polega redukcja (natury relacji). Jeśli chodzi o różnicę (1), to w modelu mechanistycznym redukcja nie zachodzi między teoriami specyfikującymi prawa naukowe, lecz między zdolnościami przypisywanymi całemu mechanizmowi (czy też systemowi zawierającemu ten mechanizm) a operacjami zorganizowanych komponentów mechanizmu. Tym samym redukcja „mechanistyczna" ma charakter *ontologiczny* (por. przypis 21), a nie reprezentacyjny, jak to jest w przypadku redukcji interteoretycznej. Jeśli zaś chodzi o różnicę (2) – dotyczącą postrzegania samej redukcji – to sprowadza się ona do faktu, iż w mechanicyzmie nie jest stawiany mocny warunek, aby redukcja przyjmowała postać wnioskowania logicznego. Dysponowanie wyjaśnieniem zdolności w kategoriach stojącego u jej podstaw mechanizmu właściwie nigdy nie oznacza, że jesteśmy w stanie logicznie wywieść zdolność z opisu części skła-

[21] Mówiąc ogólnie – w redukcjach reprezentacyjnych – redukcja zachodzi między reprezentacjami naukowymi, na przykład teoriami, modelami czy pojęciami. Od redukcji tego typu należy odróżnić redukcje ontologiczne, zachodzące między bytami określonego rodzaju, na przykład własnościami, zdarzeniami czy procesami. Rozróżnienie to zostało zaproponowane przez Roberta van Gulicka (2008).

dowych mechanizmu. W mechanistycznej koncepcji redukcji nie ma też mowy o (obecnym w modelu redukcji interteoretycznej) warunku definiowalności predykatów wyższego poziomu za pomocą predykatów niższego poziomu. Kryterium logicznej wywodliwości jest według zwolenników koncepcji mechanistycznej za mocne jako *norma* dla dobrego wyjaśnienia redukcyjnego, a jednocześnie kompletnie nieprzydatne jako narzędzie *opisu* rzeczywistych przykładów udanych redukcji dokonanych w ramach tych nauk, w których zjawiska wyjaśnia się za pomocą mechanizmów (Godfrey-Smith 2005; Craver 2007: 228–271).

Na tym etapie należy jednak wyraźnie zaznaczyć: błędem byłoby przyjęcie, że mechanicyzm to koncepcja charakteryzująca wyjaśnianie naukowe jako procedurę jednoznacznie i w całości redukcyjną. Możliwość redukowania zjawisk do ich mechanizmów ma swoje granice. Chociaż wyjaśnianie mechanistyczne jest procedurą redukcyjną, to w literaturze powszechnie rozpoznaje się fakt, że strategia mechanistyczna nie daje *kompletnego* zrozumienia zjawisk wyróżnianych na wyższych poziomach organizacji (Craver 2007: 228–271; Bechtel 2008: 142–157, 2009). Mechanistyczny model wyjaśniania pozwala nie tylko zauważyć eksplanacyjną wartość strukturalnej i funkcjonalnej dekompozycji badanego systemu, ale także eksplanacyjne ograniczenia takiej procedury. Proponuję wyróżnić cztery takie ograniczenia.

a) Ograniczenie związane z nieagregatywnością mechanizmów

Pierwsze ograniczenie wyjaśniania mechanistycznego jako strategii redukcyjnej jest związane z nieagregatywną naturą mechanizmów (Craver 2007: 135–136; Bechtel 2008: 150–151; 2009). Mechanizmy to układy działających komponentów, które odpowiadają za tę czy inną zdolność przypisywaną danemu systemowi jako całości. Co jednak istotne, wyjaśniana zdolność nie stanowi po prostu sumy operacji wykonywanych przez komponenty mechanizmu. Mechanizmy nie spełniają bowiem wyróżnionych przez Williama Wimsatta (2006a) kryteriów bycia systemem czy układem agregatywnym. Systemy czy układy agregatywne w sensie Wimsatta – takie, w których własność całości *stanowi* sumę własności części – to, po pierwsze,

systemy, w których własność systemowa (posiadana przez system jako całość) pozostaje niezmienna, gdy części systemu podlegają wzajemnej zamianie oraz gdy jest zmieniane ich rozmieszczenie. Po drugie, w systemach agregatywnych własność systemu jako całości pozostaje jakościowo identyczna[22]. pomimo tego że części są z niej eliminowane lub dodawane do niej. Po trzecie, rozłożenie systemów agregatywnych na części i ich ponowne złożenie nie skutkuje zmianą własności systemowej. Wreszcie po czwarte, w systemach agregatywnych nie dochodzi do interakcji między częściami: nie kooperują one ze sobą ani nie hamują się wzajemnie.

Mechanizmy, za pomocą których naukowcy wyjaśniają zjawiska, nie spełniają żadnego z wymienionych wyżej warunków; innymi słowy – są one *nieagregatywne*. Zmiany w rozmieszczeniu komponentów mechanizmu, wzajemne ich zamienianie, odejmowanie i dodawanie komponentów, rozkładanie ich i ponowne składanie – wszystko to skutkuje zmianą (lub całkowitą eliminacją) zdolności systemowej, za którą odpowiada działanie mechanizmu. Nieagregatywna natura mechanizmów wynika zaś z faktu, że cechują się one zawsze pewnego rodzaju przestrzenną, temporalną i aktywną organizacją. Komponenty mechanizmów nie występują po prostu w przestrzennej bliskości, lecz interreagują ze sobą w ramach pewnego przestrzennego i temporalnego porządku, a efektem tych interakcji jest własność (zdolność) wyróżniana na poziomie systemu czy mechanizmu jako całości. Każde zaburzenie wzorów interakcji między komponentami bądź zaburzenie przestrzennych czy temporalnych ram, w których te interakcje zachodzą, będzie skutkować w sposób nieuniknione zmianą własności systemowej.

W jaki sposób fakt, że mechanizmy nie są agregatami, ogranicza jednak możliwość mechanistycznej redukcji zjawisk? Jak się wydaje, nieagregatywność wprowadza dwa takie ograniczenia (por. Bechtel 2008: 150–151; 2009). Po pierwsze, nieagregatywność mechanizmów oznacza, że sama dekompozycja strukturalna i funk-

[22] Niekoniecznie pozostaje ona jednak ilościowo identyczna, na przykład kiedy odejmowanie lub dodawanie ziarenek do szklanki wypełnionej piaskiem (systemu agregatywnego) zmniejsza lub zwiększa jej masę.

cjonalna systemu nie wystarczy do wyjaśnienia żadnego zjawiska. Musi ona być każdorazowo uzupełniona wiedzą o organizacji działań i komponentów oraz o tym, w jaki sposób właśnie taka organizacja sprawia, że dany mechanizm odpowiada za wyjaśniane zjawisko. W niezwykle złożonych mechanizmach (między innymi takich, które opisuje współczesna neuronauka), cechujących się nieliniową organizacją opartą na wielokrotnych sprzężeniach zwrotnych, to właśnie wiedza o organizacji komponentów i działań będzie nie tylko kluczowa do zrozumienia zjawisk, ale też wyjątkowo trudna do pozyskania. Jak to ujmuje Bechtel (2009), wyjaśnienie mechanistyczne wymaga nie tylko spojrzenia „w dół" na składowe mechanizmu, ale także rozejrzenia się „wokół", w celu odkrycia, jak te składowe są zorganizowane.

Po drugie, można powiedzieć, że w systemach nieagregatywnych całość jest dosłownie „czymś więcej" niż tylko sumą części. Własność systemowa jest tu jakościowo różna od własności przypisywanych pojedynczym komponentom. Na przykład zdolność widzenia przypisujemy poprawnie *systemom poznawczym* (organizmom), a nie ich *komponentom*. Nie istnieje komponent systemu wzrokowego, który zajmuje się po prostu „widzeniem"; widzenie to własność systemu poznawczego jako całości, która nie może być poprawnie przypisana żadnemu z komponentów mechanizmu widzenia. Opisywanie wyodrębnionych komponentów za pomocą kategorii charakteryzujących system jako całość stanowi na ogół błąd, ponieważ funkcje czy własności systemowe nie są nigdy w ten sposób „zlokalizowane" (Wimsatt 2006a)[23]. Jak to ujmuje Craver (2007: 211––227), systemy mają zdolności czy własności przyczynowe, których nie posiadają ich komponenty. Wzięte jako odpowiednio zorganizowana całość, komponenty mogą „robić" rzeczy, których nie mogą one wykonywać wzięte z osobna[24]. Fakt ten dla niektórych zwolenni-

[23] Co nie oznacza, że tego rodzaju podejście „lokalizujące" własności systemowe w komponentach systemu nie może stanowić dobrego narzędzia heurystycznego w procesie odkrywania mechanizmów (por. wspomniana wcześniej heurystyczna teoria identyczności).

[24] Zauważmy, że w tak rozumianej tezie, iż całości posiadają własności przyczynowe różne od własności przyczynowych ich (wziętych z osobna) części, nie ma,

ków mechanicyzmu nie tylko wyklucza możliwość całkowitej redukcji własności systemowych, ale wręcz sprawia, że uzasadnione staje się mówienie o własnościach systemowych jako *emergentnych* (por.: Craver 2007: 216–217; Bechtel 2008: 129).

b) Ograniczenie związane z rolą eksplanacyjną środowiska zewnętrznego względem mechanizmu

Drugie ograniczenie mechanistycznej redukcji wiąże się z rolą eksplanacyjną środowiska zewnętrznego względem mechanizmu czy systemu, który zawiera ten mechanizm. Systemy fizyczne charakteryzują się określonymi regularnościami w zachowaniu nie tylko ze względu na ich wewnętrzną budowę, ale też ze względu na własności środowiska, w którym się one znajdują i z którym wchodzą w interakcje. Jeśli przyjrzymy się choćby samej kognitywistyce, okazuje się, że środowisko ma wieloraki wpływ na działanie mechanizmów odpowiedzialnych za zdolności poznawcze (Bechtel 2008: 152–153, 2009). Po pierwsze, wiele mechanizmów poznawczych zajmuje się regulowaniem zachowania organizmu względem okoliczności środowiskowych, w związku z czym ich działanie jest często „czułe" na te okoliczności. Niejednokrotnie „ciężar" wyjaśnienia zachowania systemu jako całości będzie więc spoczywał na wskazaniu warunków środowiskowych, w jakich on się znajduje, a nie na samej jego wewnętrznej organizacji. Na przykład mechanizmy poznawcze, których działanie jest oparte na uczeniu w wyniku interakcji ze środowiskiem, będą działały różnie w zależności od kształtu środowiska rozwojowego, w którym znalazł się system poznawczy. Działanie tych mechanizmów będzie w zasadniczy sposób zależne od tego, czy w środowisku występują przewidywalne regularności w zachowaniu

jak się wydaje, nic metafizycznie problematycznego. Teza ta staje się problematyczna dopiero wtedy, gdy mowa nie o poziomach kompozycji, lecz – realizacji, gdzie funkcjonalna własność wyższego rzędu miałaby mieć moce przyczynowe różne od realizujących ją własności fizycznych (Craver 2007: 211––227). Poziomy kompozycyjnej organizacji mechanizmów nie są jednak poziomami realizacji (por. omówienie poziomów mechanizmów zawarte w sekcji poprzedniej oraz omówienie związku między mechanicyzmem a wieloraką realizacją zawarte w sekcji 2.3.2).

obiektów czy partnerów interakcji społecznych. Wyjaśnianie określonych deficytów poznawczych może więc czasem wymagać nie tylko wskazania zaburzeń na poziomie wewnętrznej, mechanistycznej organizacji, ale także odwołania się do określonych faktów na temat środowiska rozwojowego, w jakim znalazł się system jako całość.

Po drugie, wielu współczesnych badaczy argumentuje, że elementy środowiska zewnętrznego często aktywnie uzupełniają działanie wewnętrznych mechanizmów poznawczych (por.: Kirsh, Maglio 1994; Hutchins 1995; Clark, Chalmers 2008; Robbins, Aydede 2008). Ludzie wykonujący niektóre zadania wydają się „odciążać" obliczeniowo własny układ nerwowy przez odpowiednie manipulowanie elementami środowiska. Różnego rodzaju artefakty i technologie pełnią funkcję „rusztowań", które kompensują ograniczenia wewnętrznych mechanizmów i dzięki którym mogą być realizowane cele niedostępne „za pomocą" samych mechanizmów wewnętrznych[25]. Bez powołania się na kluczową rolę takich rusztowań, pewne interesujące kognitywistów zjawiska pozostaną bez wyjaśnienia. Powracając zatem do metafory Bechtela (2009), wyjaśnienie określonego zjawiska wymaga często nie tylko spojrzenia „w dół" (na komponenty i operacje składowe mechanizmu) i „wokół" (na organizację komponentów i działań składowych), ale także „w górę" – na to, jak mechanizm czy system jako całość jest usytuowany w ramach określonych warunków środowiskowych.

[25] Można tu postawić pytanie, czy tego rodzaju zewnętrzne zasoby uzupełniają albo wspomagają system poznawczy, czy może raczej stają się jego *częściami*, za którą to możliwością optują zwolennicy koncepcji rozszerzonego umysłu (por. Clark, Chalmers 2008). Przyjmując to drugie stanowisko, należałoby uznać, że skóra organizmu nie wyznacza granicy systemu poznawczego. W takim wypadku powołanie się na eksplanacyjną rolę środowiska nie oznaczałoby jeszcze, że wychodzimy poza granice wyjaśniania mechanistycznego. Przy takim scenariuszu nasze wyjaśnienia nadal powoływałyby się na mechanizmy (zorganizowane układy aktywnych komponentów systemu poznawczego), nawet jeśli byłyby to mechanizmy znajdujące się poza granicą wyznaczaną przez skórę organizmu. Aby zablokować taką możliwość, moglibyśmy przyjąć rozumienie systemu poznawczego, zgodnie z którym jest on kontekstowo stałym czy niezmiennym układem mechanizmów poznawczych (por. Rupert 2009). Zagadnienie to wykracza jednak znacznie poza cele tej pracy, może ono więc być tu jedynie zaznaczone, a nie szerzej podjęte.

c) Ograniczenie związane z pluralizmem eksplanacyjnym

Trzecie ograniczenie mechanistycznej redukcji zjawisk polega na tym, że wiele zjawisk domaga się wyjaśnienia za pomocą strategii innych niż (jedynie) mechanistyczna. W zgodzie z przyjętym tu pluralizmem eksplanacyjnym często może zdarzać się, że specyficznie mechanistyczne pytanie „jak?" nie jest jedynym, na które trzeba znać odpowiedź, aby można było uznać, że dysponujemy kompletnym wyjaśnieniem danego eksplanandum. Oprócz pytania „jak?" czasem można bowiem także zapytać o to, „dlaczego" dany system dysponuje daną zdolnością. To drugie pytanie można na przykład rozumieć jako pytanie o *etiologię przyczynową* zdolności lub jako pytanie o *teleologiczne* wyjaśnienie tej zdolności (de Pinedo, Noble 2008; por. też Nikolasa Tinbergena odróżnienie na „bliższe" i „dalsze" wyjaśnienia zjawisk biologicznych: Tinbergen 1963). Przyjrzyjmy się tym ostatnim, teleologicznym rodzajom wyjaśnień. Będą one miały identyczne eksplanandum, co pewne odpowiadające im wyjaśnienia mechanistyczne. Jednak użyta w nich strategia eksplanacyjna będzie się zasadniczo różnić od strategii opartej na wyjaśnianiu za pomocą mechanizmu. Weźmy pod uwagę przykład z zakresu nauk kognitywnych, mianowicie wyjaśnianie zdolności do czytania umysłów (posługiwania się psychologią potoczną). Możemy rzecz jasna poszukiwać mechanistycznego wyjaśnienia tej zdolności, odpowiadającego na pytanie o to, *jak* istoty ludzkie – oraz, ewentualnie, inne naczelne – przypisują sobie oraz innym stany mentalne. Moglibyśmy na przykład postulować, że mechanizm czytania umysłów wykorzystuje ukrytą teorię (por. Carruthers 2009), albo że jest oparty na symulacji własnych lub cudzych procesów mentalnych (por. Goldman 2006). Nawet jednak odkrywszy ten mechanizm, nadal można będzie zapytać o to, *dlaczego* ludzie posiadają zdolność do przypisywania sobie i innym stanów mentalnych. Naturalistyczna odpowiedź na to pytanie może odwołyać się chociażby do tezy, że zdolność czytania umysłów jest adaptacją do określonych warunków, w których ewoluował umysł *Homo sapiens* i jego przodków. Na przykład ta kompetencja może być odpowiedzią na wymogi związane ze współzawodnictwem (por. Whiten 1997) lub kooperacją (por. Sterelny 2003: 123–145) między jednostkami w ra-

mach życia społecznego wczesnych hominidów. Choć oba wyjaśnienia – mechanistyczne i adaptacyjne/teleologiczne – mają jeden przedmiot, to wyjaśniają go w różny sposób. To drugie wyjaśnienie, ale już nie pierwsze, będzie się z konieczności odwoływać do pewnych historyczno-środowiskowych faktów dotyczących wyjaśnianego zjawiska. Co więcej, poprawność tego drugiego wyjaśniania pozostaje przynajmniej *prima facie* niezależna od tego, które wyjaśnienie mechanistyczne jest poprawne. To, do czego zdolność czytania umysłów stanowi adaptację (zakładając, że w ogóle stanowi), jest niezależne od tego, czy u jego podstaw leży mechanizm (*quasi*-)teoretycznego wnioskowania, czy też mechanizm symulacyjny. Uogólniając, kompletne naukowe wyjaśnienie wielu – być może większości – zjawisk będzie wymagało czegoś więcej, niż jedynie wskazania stojących u ich podstaw mechanizmów.

d) Ograniczenie związane z wielopoziomową naturą wyjaśnień mechanistycznych

Czwarte wreszcie ograniczenie mechanistycznej redukcji zjawisk jest związane z faktem, iż wyjaśnienia mechanistyczne są w nieunikniony sposób wielopoziomowe. Redukcjonizm bywa ściśle wiązany z ideą, że wyjaśnienia, teorie czy modele z niższego poziomu będą w stanie ostatecznie zastąpić wyjaśnienia, teorie czy modele z wyższego poziomu (por. np. Bickle 2003). Udana redukcja powinna skutkować sytuacją, w której cała wymagana „praca eksplanacyjna" zostaje wykonana przez koncepcję (teorię, model, wyjaśnienie) sformułowaną na najniższym poziomie, co czyni koncepcje sformułowane na poziomach wyższych zbędnymi lub przejawiającymi jedynie heurystyczną wartość. Zwolennicy mechanistycznego podejścia do natury wyjaśniania powszechnie odrzucają tego rodzaju optykę (Bechtel, Abrahamsen 2005; Craver 2007: 256–266; Bechtel 2008: 155––157, 2009; Machamer, Darden, Craver 2011; por. też: Gold, Stoljar 1999; Miłkowski, Poczobut 2005; Poczobut 2009 – autorzy ci bronią podobnej pozycji, nie przyjmując *explicite* założeń mechanicyzmu). Zwracają oni uwagę na fakt, że opisy i wyjaśnienia badanego systemu sformułowane na kolejnych „piętrach" jego mechanistycznej organizacji nie są w praktyce zastępowane tymi sformułowanymi na

piętrze „najniższym"; wszystkie poziomy są bowiem *niezbywalne*, jeśli poszukujemy kompletnego wyjaśnienia danego zjawiska.

Zwróćmy uwagę, że mowa teraz o poziomach *opisu i wyjaśniania*, a nie o metafizycznych poziomach mechanistycznej *organizacji*. Te dwie kategorie są jednak ze sobą w mechanicyzmie powiązane. Jak zauważa Bechtel, z perspektywy modelu mechanistycznego każdy poziom organizacji „szkatułkowego", hierarchicznego układu mechanizmów wyznacza jeden poziom opisu i wyjaśniania (poziom w sensie epistemicznym; por. Bechtel 1994). Na przykład każdemu poziomowi organizacji w hierarchii mechanizmów odpowiedzialnych za pamięć przestrzenną szczurów (por. sekcja 2.1.1 i rysunek 2) odpowiada pewien poziom wyjaśniania. Na każdym z tych poziomów wyjaśniania badacze posługują się osobnym zestawem predykatów, za pomocą którego nie można opisywać komponentów z niższego poziomu. Szczur i jego relacje ze środowiskiem są charakteryzowane za pomocą innych predykatów niż proces formowania się długotrwałych wzmocnień synaptycznych. Na każdym z poziomów badacze formułują generalizacje dotyczące pewnych ogólnych zależności między komponentami, które nie stosują się do komponentów na niższych poziomach. Generalizacje dotyczące sposobów poruszania się szczurów w labiryntach nie mogą być orzekane o aktywności receptorów NMDA. Wreszcie na każdym z poziomów badacze posługują się innymi technikami badawczymi czy eksperymentalnymi. Zgodnie z koncepcją mechanistyczną ten eksplanacyjny, konceptualny i metodologiczny pluralizm – zawsze towarzyszący wyjaśnieniom mechanistycznym – jest naturalny i nieunikniony; nie stanowi jedynie przejściowej sytuacji, której punktem dojścia będzie zastąpienie wszelkich wyjaśnień sformułowanych na wyższych poziomach przez wyjaśnienie sformułowane na poziomie molekularnym (Craver 2007: 256–266). Wyższe poziomy organizacji są postrzegane jako *realne* – choćby dlatego, że mają własności przyczynowe, których nie posiadają z osobna komponenty mechanizmu niższego poziomu (por. Hacking 1983). Jednocześnie wyjaśnienia i opisy skupiające się na wyższych poziomach są też *epistemicznie niezbywalne* w procesie wyjaśniania danego zjawiska – choćby dlatego, że poszukiwania mechanizmów z niższego poziomu są w prak-

tyce zależne od technik badawczych wymagających interwencji eksperymentalnych na poziomie wyższym (por. opisaną wcześniej rolę eksperymentów odgórnych w odkrywaniu mechanizmów). Badania prowadzone na różnych poziomach mechanistycznej organizacji nie zastępują i nie eliminują się wzajemnie, lecz koewoluują ze sobą (Craver 2007: 256–266). Jak stwierdza Craver (2007: 228–271), normatywny ideał przyświecający wyjaśnieniom mechanistycznym nie polega na zastępowaniu jednych poziomów wyjaśniania innymi, lecz na *integrowaniu* wyników badań prowadzonych na różnych poziomach organizacji (na przykład od molekularnego po behawioralny), tak, by można było zobaczyć, że badania te zajmują się *de facto* jednym układem hierarchicznie zorganizowanych mechanizmów (na przykład tych stojących u podstaw pamięci przestrzennej).

Jak wynika z powyższych rozważań, redukowanie zjawisk do ich mechanizmów ma swoje granice. Wspomniałem już wcześniej, iż niektórzy autorzy twierdzą wręcz, że z perspektywy mechanicyzmu daje się uzgodnić redukcjonizm z emergentyzmem. Jak stwierdza Bechtel:

> Termin „emergencja" jest używany w wielu znaczeniach, czasem takich, które tajemniczo implikują, iż niekiedy do Wszechświata są dodawane pewne radykalnie nowe byty. Emergentyzm rozumiany tak, jak ja stosuję tu ten termin, rozpoznaje po prostu fakt, że cały system działa w sposób, który wykracza poza działania jego części. Uderza to do pewnego stopnia w skrajny redukcjonizm, w tym sensie, że zachowanie całego systemu musi być badane na jego własnym poziomie i za pomocą stosownych na tym poziomie narzędzi. Badania prowadzone na poziomie systemu jako całości dysponują pewnym stopniem niezależności [...]. Uznanie takiej niezależności nie wymaga jednakże przyjęcia żadnych metafizycznie tajemniczych bytów postulowanych, takich jak siły życiowe. Co więcej, jest ono [badanie całości – P. G.] nie tylko spójne z badaniami komponentów i ich operacji, ale przez te badania ograniczane [*constrained*]. (I na odwrót, badania części i ich operacji są ograniczane przez badania mechanizmu jako całości) (Bechtel 2008: 129).

Mechanicyzm wydaje się więc „przyjazny" zarówno redukcjonizmowi, jak i przynajmniej niektórym umiarkowanym sformułowaniom emergentyzmu, które unikają mówienia o „radykalnie nowych" bytach. Mechanistyczny model wyjaśniania jest zatem zbieżny z twierdzeniami filozoficznymi głoszącymi zgodność (pewnych form) redukcjonizmu z (pewnymi formami) emergentyzmu (por. np.: Wimsatt 2006a; Poczobut 2009). Pomijając jednak kwestię emergencji, wystarczy podkreślić, że wszystkie powyższe obserwacje prowadzą do wniosku, iż mechanistyczny model wyjaśniania nie poddaje się jednoznacznemu określeniu w ramach opozycji „redukcjonizm–antyredukcjonizm". Z jednej strony postuluje on sprowadzanie zachowań systemów/mechanizmów jako całości do działań ich komponentów przez dekompozycję strukturalną i funkcjonalną, co stanowi bez wątpienia procedurę nietrywialnie redukcyjną. Z drugiej strony z punktu widzenia mechanicyzmu jest dość oczywiste, że taka strategia nie generuje wyjaśnień kompletnych. Aby je bowiem uzyskać, potrzeba dodatkowo wiedzy o organizacji komponentów i ich działań, jak również o osadzeniu systemu jako całości w środowisku. Niejednokrotnie wskazanie mechanizmu nie jest także jedyną strategią eksplanacyjną, za pomocą której możemy wyjaśnić dane zjawisko. Wreszcie wyjaśnianie mechanistyczne w nieunikniony sposób odwołuje się do wielu poziomów organizacji i wyjaśniania: wielopoziomowość jest tu normą, a nie niezręcznym stanem przejściowym.

2.3. Aplikacja modelu mechanistycznego do wyjaśniania w kognitywistyce

2.3.1. Wyjaśnianie mechanistyczne w kognitywistyce: zasadnicze założenia

Zgodnie z centralnym założeniem stojącym u podstaw rozważań prezentowanych w następnych rozdziałach, wyjaśnienia zdolności poznawczych w naukach kognitywnych mają w ogromnej mierze charakter mechanistyczny. Celem kognitywistyki jest odkrycie

i opisanie wewnętrznej, funkcjonalnej „architektury" określonej klasy systemów fizycznych, mianowicie systemów mających zdolności poznawcze. Pytania kognitywistyki to w dużym stopniu pytania o to, „jak" systemy poznawcze wykonują czynności, w których manifestują się posiadane przez nie zdolności. Udzielenie odpowiedzi na te pytania wymaga wskazania odpowiednich mechanizmów. Truizm, że celem kognitywistyki jest odkrywanie „mechanizmów poznania", należy zatem rozumieć całkiem dosłownie. Mechanicyzm dookreśla jedynie, jak powinno się rozumieć owe „mechanizmy" oraz na czym polega wyjaśnianie zjawisk za ich pomocą.

Założenie o roli wyjaśnień mechanistycznych w kognitywistyce wykorzystam dalej w celu sformułowania koncepcji wyjaśnień reprezentacyjnych. Będę przyjmował, że wyjaśnienia zjawisk poznawczych za pomocą *reprezentacji mentalnych* stanowią formę wyjaśnień mechanistycznych. Częste sformułowania o „wewnętrznych" reprezentacjach powinny być rozumiane dosłownie. Chodzi tu ostatecznie o wewnętrzne, przyczynowo aktywne struktury systemu poznawczego (zawarte w ośrodkowym układzie nerwowym), które wchodzą w skład mechanizmów wyjaśniających zdolności posiadane przez ten system. Innymi słowy, będę przyjmował, że reprezentacje wyjaśniają zjawiska, o ile w systemie poznawczym istnieją wewnętrzne mechanizmy, które z reprezentacji korzystają. Następny rozdział będzie poświęcony nadaniu określonego sensu temu twierdzeniu. W chwili obecnej skupmy się jeszcze na związku między mechanicyzmem a kognitywistyką jako taką. Zacznijmy od wprowadzenia trzech ważnych uzupełnień do twierdzenia o wyjaśnianiu w kognitywistyce jako wyjaśnianiu mechanistycznym.

Po pierwsze, nawet jeśli wyjaśnianie przez wskazywanie mechanizmów zjawisk odgrywa zasadniczą rolę w kognitywistyce, to nie wynika z tego, że ogół, a nawet większość *praktyki badawczej* kognitywistów koncentruje się na poszukiwaniu wyjaśnień mechanistycznych. W istocie znaczna część odkryć eksperymentalnych w kognitywistyce dostarcza informacji o tym, co Cummins (2000) nazywa „efektami", czyli o pewnych ogólnych zależnościach czy regularnościach cechujących procesy poznawcze; regularnościach dotyczących choćby interakcji między modalnościami zmysłowy-

mi, percepcją a pojęciami, wiekiem a rozwojem określonych zdolności poznawczych, między obciążeniem pamięci roboczej a efektywnością w realizowaniu określonych zadań eksperymentalnych i tak dalej. Choć przyjmowane tu przeze mnie stanowisko nie neguje takiego stanu rzeczy, to implikuje ono także, że odkrycia tego rodzaju ogólnych zależności *nie* mają jeszcze wartości eksplanacyjnej, przynajmniej nie jako wyjaśnienia *mechanistyczne*. Za Cumminsem przyjmuję zatem, że tego rodzaju badania pozwalają na sformułowanie eksplanandów dla wyjaśnień mechanistycznych – możemy pytać o mechanizm wyjaśniający, na przykład, odkryte wzory interakcji między modalnościami zmysłowymi – jednak nie dostarczają one (mechanistycznych) eksplanansów.

Po drugie, w zgodzie z ideą pluralizmu eksplanacyjnego, nie przyjmuję w tej pracy, że *wszelkie* wyjaśnienia w kognitywistyce są mechanistyczne. Nawet jeśli znacząca czy zasadnicza część z nich opiera się na opisywaniu mechanizmów, dopuszczam możliwość istnienia wyjaśnień stosujących strategie inne niż mechanistyczna. Zgodnie z poczynionymi tu już ustaleniami takie strategie mogą opierać się na wyjaśnianiu zjawisk przez odwołanie między innymi (1) do etiologii przyczynowej na tym samym poziomie organizacji (Craver 2007: 74, 93–104) czy (2) do określonych presji selekcyjnych i funkcji adaptacyjnych, jak chociażby w psychologii ewolucyjnej (por. Barkow, Cosmides, Tooby 1992). Inaczej rzecz ujmując, choć zasadniczy cel nauk kognitywnych to udzielanie odpowiedzi na pytania „jak?" (za pomocą wskazywania stosownych mechanizmów poznania), z pewnością byłoby fałszywe twierdzenie, że kognitywiści nie poszukują także (przyczynowo-etiologicznych czy ewolucyjno-teleologicznych) odpowiedzi na pytania „dlaczego?".

Oczywiście istnieją granice pluralizmu eksplanacyjnego w kognitywistyce. Nie jest tak, że wszelkie rodzaje wyjaśniania naukowego są równie rozpowszechnione w naukach kognitywnych, jak wyjaśnienia mechanistyczne. Jak już wspomniałem, rola wyjaśnień odwołujących się do praw pozostaje w kognitywistyce marginalna (zakładając, że takie wyjasnienia w ogóle występują). Ponadto stojąc na stanowisku mechanicyzmu, trzeba odrzucić te koncepcje wyjaśniania, które są z koncepcją mechanistyczną w oczywisty

sposób niespójne. Dlatego też nie do przyjęcia jest „bezlitosny redukcjonizm" (*ruthless reductionism*) Johna Bickle'a (2003), zgodnie z którym kompletne wyjaśnienia zjawisk poznawczych mogą zostać sformułowane na najniższym (molekularnym) poziomie organizacji systemu poznawczego. Jak już zaznaczyłem w poprzedniej sekcji, rzeczywiste wyjaśnienia mechanistyczne odwołują się do różnych poziomów organizacji. Założenie o istnieniu tych poziomów wydaje się nieuniknione i nieusuwalne w realnej praktyce badawczej kognitywistów. Nie istnieje jeden, najniższy, czysto molekularny lub biologiczny poziom, na którym można by formułować kompletne wyjaśnienia zdolności poznawczych (por.: Gold, Stoljar 1999; Looren de Jong, Schouten 2005; Miłkowski, Poczobut 2005; Craver 2007: 256––266; Poczobut 2009).

Po trzecie, warto zapytać, czy istnieje coś, co stanowiłoby o specyfice mechanistycznych wyjaśnień w kognitywistyce i odróżniało je od mechanistycznych wyjaśnień stosowanych w innych dyscyplinach? Jak się wydaje, na pytanie to można odpowiedzieć na dwa sposoby. Pierwszy rodzaj odpowiedzi jest liberalny. Zgodnie z tym podejściem jedynym, co wyróżnia wyjaśnienia mechanistyczne w kognitywistyce, są ich eksplananda. Eksplanandami tymi są funkcje czy zdolności umysłowe (poznawcze), takie jak między innymi formowanie pojęć, kategoryzacja percepcyjna, percepcja słuchowa, podejmowanie decyzji, nabywanie języka ojczystego czy integracja sensomotoryczna. Drugi rodzaj odpowiedzi na pytanie o specyfikę wyjaśnień mechanistycznych w naukach kognitywnych jest bardziej restrykcyjny, ponieważ specyfikuje on nie tylko naturę kognitywistycznych eksplanandów, ale też eksplanansów. W takim przypadku postuluje się, że kognitywistyka wyjaśnia interesujące ją zjawiska za pomocą specyficznego rodzaju mechanizmów, które w jakimś sensie różnią się od mechanizmów odkrywanych przez przedstawicieli innych dyscyplin (na przykład działają one na podstawie reprezentacji lub są obliczeniowe).

W ramach prezentowanej pracy będę przyjmował to pierwsze, liberalne stanowisko w sprawie specyfiki mechanistycznych wyjaśnień w kognitywistyce. Chcę być możliwie neutralny czy inkluzywny, jeśli chodzi o to, jaka może być natura mechanizmów stojących

u podstaw zdolności poznawczych. Będę tu rzecz jasna bronił możliwości uzgodnienia mechanicyzmu z tezą o roli eksplanacyjnej reprezentacji mentalnych. Jednak zagadnienie, czy (oraz jak) reprezentacje mogą spełniać rolę eksplanacyjną w kognitywistyce, jest konceptualnie niezależne od kwestii dotyczącej tego, czy reprezentacje stanowią narzędzie eksplanacyjne specyficzne dla kognitywistyki. Przyjęcie tej ostatniej tezy za prawdziwą implikowałoby przecież, że wszelkie mechanizmy poznawcze, i tylko one, posługują się reprezentacjami. Przyjęcie takiego twierdzenia byłoby jednak nieuzasadnione i zasadniczo nieuczciwe dla antyreprezentacjonizmu. W ramach prowadzonych tu rozważań dopuszczam możliwość, że przynajmniej część mechanizmów odpowiadających za zjawiska poznawcze nie posługuje się reprezentacjami.

2.3.2. Mechanicyzm, wyjaśnienia funkcjonalne i wieloraka realizowalność

Jak się jednak okazuje, mechanistyczny model wyjaśniania nie może zostać zastosowany do kognitywistyki w sposób całkowicie bezproblemowy. Wśród dość szeroko akceptowanych w kognitywistyce i filozofii umysłu – przynajmniej w tradycji analitycznej – założeń dotyczących natury wyjaśniania w naukach o poznaniu znajduje się także takie, które wydaje się *prima facie* nie do pogodzenia z mechanicyzmem. Chodzi mianowicie o twierdzenie, zgodnie z którym wyjaśnianie w kognitywistyce ma charakter *funkcjonalny*, przy czym jedną z jego form jest wyjaśnianie obliczeniowe.

Dlaczego twierdzenie o funkcjonalnej naturze wyjaśniania w naukach kognitywnych stanowi problem z punktu widzenia mechanicyzmu? Zauważmy, że mechanizmy są bytami fizycznymi. Mechanizmy poznawcze to w takiej perspektywie mechanizmy neurobiologiczne: funkcjonalnie scharakteryzowane, zorganizowane komponenty (na przykład grupy czy populacje neuronów) ośrodkowego układu nerwowego. Budowanie mechanistycznych wyjaśnień zjawisk poznawczych wymaga, by postulowane mechanizmy były neurobiologicznie adekwatne, czyli zgodne z tym, jak działają realne, biologiczne mózgi. Z punktu widzenia mechanicyzmu, neu-

ronauka jest więc integralną i posiadającą centralne znaczenie częścią nauk kognitywnych. Tymczasem przyjęcie funkcjonalistycznego spojrzenia na naturę wyjaśniania w kognitywistyce wydaje się stać w sprzeczności z takim twierdzeniem. Dostarczanie wyjaśnienia pewnego zjawiska poznawczego w kategoriach funkcjonalnych zdaje się niezależne względem „strukturalnych" faktów dotyczących budowy i organizacji mózgu, a także – przez te fakty nieograniczane. Można na przykład twierdzić, że co prawda poznanie szczegółów dotyczących neuronalnej implementacji pewnego modelu obliczeniowego będzie naukowo przydatne, jednak model ten posiada wartość eksplanacyjną niezależnie od tego, jak jest on zaimplementowany w mózgu. Z takiego punktu widzenia kognitywistyka pozostaje autonomiczna względem neuronauki w następującym znaczeniu: (1) wyjaśnienia nie są tu ograniczane i weryfikowane przez fakty neurobiologiczne; (2) kognitywiści mogą formułować kompletne i poprawne wyjaśnienia zjawisk niezależnie od tego, jakie są fakty dotyczące organizacji i sposobu działania mózgu (Piccinini, Craver 2011).

Wydaje się jednak, że podejścia funkcjonalne i mechanistyczne mogą zostać uzgodnione, a niekompatybilna z mechanicyzmem teza o autonomii kognitywistyki względem neuronauki – odrzucona. Gualtiero Piccinini i Craver (2011) wyróżniają trzy typy wyjaśnień funkcjonalnych stosowanych przez kognitywistów[26]: (1) analizę zadań, w której funkcjonalnie dekomponuje się pewną zdolność na szereg zdolności składowych (na przykład dekompozycja pamięci semantycznej na kodowanie, przechowywanie i odzyskiwanie); (2) wyjaśnianie zdolności poznawczej przez postulowanie

[26] Piccinini oraz Craver (2011) piszą co prawda o autonomii psychologii (a nie kognitywistyki jako takiej) względem neuronauki, jednak wydaje się, że *de facto* za pomocą terminu „psychologia" oznaczają oni dowolną dyscyplinę, która wyjaśnia zjawiska poznawcze czysto funkcjonalnie. Ich argumentacja stosuje się równie dobrze do sztucznej inteligencji, o ile ta miałaby pretensje do bycia neutralną względem neuronauki. Z tego właśnie powodu w tej pracy formułuję kwestię autonomii wyjaśnień funkcjonalnych w kategoriach relacji między neuronauką a czysto „funkcjonalistyczną" kognitywistyką, a nie jedynie psychologią czy inną dyscypliną składową kognitywistyki.

wewnętrznych, funkcjonalnie indywiduowanych stanów systemu i procesów polegających na zmianie tych stanów w czasie; (3) „pudełkowe" (*boxological*) wyjaśnienia zdolności poznawczych w kategoriach wchodzących ze sobą w interakcje komponentów systemu („czarnych skrzynek"), które to komponenty są indywiduowane czysto funkcjonalnie. Choć prowadzony przez Piccininiego i Cravera wywód jest złożony i szczegółowy, to prezentowany przez nich argument przeciwko autonomii „funkcjonalistycznej" kognitywistyki względem neuronauki okazuje się stosunkowo prosty. Dowolna zdolność poznawcza może mieć kilka alternatywnych modeli funkcjonalnych. Rzekoma poprawność każdego z nich będzie na ogół uzasadniana przez odwołanie do jego fenomenalnej adekwatności, to znaczy faktu, że model ten umożliwia przewidywanie zachowania systemu w określonych sytuacjach (przy otrzymaniu określonych danych wejściowych). Jednakże skuteczność predykcyjna nie wystarcza do wyjaśnienia zjawiska. Nie jest ona nawet w tym celu konieczna. Możemy rozumieć działanie systemu (dysponować jego wyjaśnieniem), nie potrafiąc go przewidzieć, a zarazem zdolności do przewidzenia działania systemu nie zawsze towarzyszy rozumienie, jak on działa. Jaki dodatkowy warunek musi zatem spełniać model funkcjonalny, aby zyskał status wyjaśnienia? Otóż mechanicyzm proponuje intuicyjną i wiarygodną odpowiedź na to pytanie. Zgodnie z nią kryterium eksplanacyjnej wartości dowolnego modelu funkcjonalnego stanowi możliwość uzgodnienia dekompozycji funkcjonalnej z dekompozycją *strukturalną* danego systemu poznawczego. Innymi słowy, musi istnieć możliwość przypisania funkcji specyfikowanych w takim modelu konkretnym komponentom mechanizmu, pojmowanego jako pewien realny, fizyczny (a nie jedynie abstrakcyjny) układ. Tym samym rzeczywistą wartość czy moc eksplanacyjną ma ten spośród alternatywnych modeli funkcjonalnych, który stanowi adekwatny (lub najbardziej przybliżony) opis funkcjonalnej organizacji komponentów mechanizmu rzeczywiście odpowiedzialnego za tę czy inną zdolność poznawczą. W związku z tym (1) zdolności składowe wyróżniane w ramach analizy zadań powinny być rozumiane jako zdolności przypisywane komponentom mechanizmu; (2) wewnętrzne stany modelowanego systemu

powinniśmy rozumieć jako stany, w których znajdują się jego komponenty; (3) „czarne skrzynki" w modelach „pudełkowych" powinniśmy rozumieć jako komponenty mechanizmu cechujące się określonym umiejscowieniem funkcjonalnym w ramach tego systemu. Mówiąc obrazowo: aspirujące do miana wyjaśnień modele funkcjonalne stoją przed trybunałem faktów „strukturalnych" dotyczących składu i organizacji realnych, fizycznych mechanizmów.

Wniosek, jaki wyciągają Piccinini i Craver (2011) z tych rozważań nie głosi, że budowanie czysto funkcjonalnych modeli to czynność eksplanacyjnie bezwartościowa. Chodzi im raczej o to, że praktyka ta nie jest ani odrębna, ani autonomiczna względem praktyki budowania wyjaśnień mechanistycznych. Autorzy ci formułują tę ideę, posługując się rozróżnieniem na szkice oraz schematy mechanizmów, o którym wspomniałem w sekcji 2.1.1. Szkice mechanizmów są na ogół tworzone we wczesnych stadiach badania naukowego. Stanowią one wynik prób zrozumienia budowy mechanizmu odpowiedzialnego za dane zjawisko, jednak są niekompletne, pełne uproszczeń, idealizacji, a także często wykorzystują w opisie mechanizmu nieprecyzyjne „terminy-wypełniacze" (*filler terms*). Systematyczne uzupełnianie luk obecnych w szkicu i związany z tym wzrost jego deskryptywnej adekwatności skutkuje ostatecznie zmienieniem go w schemat mechanizmu, czyli taki opis mechanizmu, który można określić, przynajmniej w przybliżeniu, jako kompletny i precyzyjny[27]. Wedle Picciniego i Cravera funkcjonalne modele zdolności poznawczych powinniśmy traktować jako szkice mechanizmów. Są one wstępnymi reprezentacjami mechanizmów, często budowanymi w sytuacji, w której nie znamy jeszcze faktów dotyczących ich realnej, fizycznej struktury. Jeśli takie czysto funkcjonalne szkice mają zamienić się w eksplanacyjnie „pełnowartościowe" schematy, muszą być one systematycznie rewidowane i uzupełniane w świetle wiedzy na temat rzeczywistej strukturalnej organizacji stosownych mechanizmów zlokalizowanych w ośrodkowym układzie nerwowym (mózgu). Tym samym modele funkcjonalne (1) podlegają ograniczeniu

[27] Jak się wydaje, nie sposób wskazać jednoznacznie „punkt", w którym szkic mechanizmu staje się schematem. Kategorie te należy uznać za rozmyte.

niom i są weryfikowane w świetle wiedzy o strukturalnej organizacji mózgu, a także (2) tylko w świetle tej wiedzy mogą one być uznawane za poprawne lub niepoprawne. Akceptacja wartości opisów czy modeli funkcjonalnych w kognitywistyce nie implikuje nieprzyjaznej mechanicyzmowi tezy o autonomii tej dyscypliny względem neuronauki. Zachodzi ciągłość między tworzonymi przez psychologów czy badaczy zajmujących się sztuczną inteligencją modelami funkcjonalnymi systemu poznawczego a neuronaukowymi odkryciami dotyczącymi budowy i organizacji mózgu. Dzieje się tak dlatego, że obie te dyscypliny są, choć w inny sposób i od innej strony, zaangażowane w realizowanie jednego celu polegającego na dostarczaniu mechanistycznych wyjaśnień zdolności poznawczych.

Jak zostało już zaznaczone, bardzo istotną podkategorię wyjaśnień funkcjonalnych w kognitywistyce stanowią wyjaśnienia obliczeniowe. Czy także do nich można zastosować powyższą argumentację? Czy i one są formą wyjaśnień mechanistycznych? Wydaje się, że przedstawiona powyżej strategia argumentacyjna wspiera twierdzącą odpowiedź na to pytanie. Możemy dysponować kilkoma różnymi, jednak za każdym razem predykcyjnie skutecznymi obliczeniowymi modelami zdolności poznawczej charakteryzującej pewien system. Czynnikiem różnicującym moc eksplanacyjną tych poszczególnych alternatywnych modeli będzie zaś to, jak dobrze opisują one, w jaki sposób dany system *rzeczywiście* generuje określone wzorce zachowań, to znaczy jakie wewnętrzne mechanizmy za to odpowiadają (Buller 1993; Piccinini 2007a, 2007b; Kaplan 2011; Miłkowski 2013).

Na czym polega jednak to, że pewien model obliczeniowy dobrze opisuje rzeczywistą przyczynową strukturę mechanizmu? To osobne zagadnienie, którego nie można tu systematycznie rozwinąć. Mówiąc jednak bardzo ogólnie, możemy potraktować model obliczeniowy jako model specyfikujący ciąg liter (symboli) w skończonym alfabecie oraz listę instrukcji (program) określających reguły przetwarzania jednych ciągów liter z tego alfabetu na inne (Piccinini 2007a, 2007b; por. Miłkowski 2013). W mechanizmie odpowiadającym temu modelowi komponenty (lub stany, w jakich się one znajdują) odpowiadają literom alfabetu. Komponenty stanowiące egzem-

plarze jednego typu litery z alfabetu będą zbliżone pod względem posiadanych własności fizycznych, a także będą różniły się fizycznie od komponentów stanowiących egzemplarze liter innego typu. Mechanizm powinien być przyczynowo „wrażliwy" na tego rodzaju podobieństwa i różnice, tak że komponenty należące do jednego typu będą w nim przetwarzane inaczej, niż komponenty podpadające pod inne typy liter. Jeśli komponenty są w ramach mechanizmu manipulowane (przetwarzane) w zgodzie z zestawem instrukcji zawartym w modelu obliczeniowym, to można uznać, że model ten poprawnie opisuje działanie tego mechanizmu. Nie oznacza to rzecz jasna, iż ten sam model obliczeniowy nie może poprawnie opisywać mechanizmów o bardzo różnych własnościach fizycznych, konkretnie mechanizmów opierających się na różnych materiałach fizycznych. Modele obliczeniowe są tym samym neutralne ze względu na nośnik czy też medium (na przykład elektryczne lub elektro-chemiczne), w którym przebiegają obliczenia. Pamiętajmy jednak, że taka neutralność nie implikuje globalnej neutralności względem mechanistycznej organizacji systemów fizycznych, których działanie jest obliczeniowo wyjaśniane. Choć modele obliczeniowe nie niosą pewnych zobowiązań dotyczących mechanizmów – a konkretnie zobowiązań odnoszących się do materiału fizycznego stanowiącego nośnik obliczeń – to jednak ich poprawność i moc eksplanacyjna zależą od innych własności tych mechanizmów, mianowicie od *funkcjonalnej organizacji ich komponentów*[28]. Raz jeszcze, dekompozycja funkcjonalna (tu: obliczeniowa) systemu zyskuje status wyjaśnienia tylko wtedy, gdy można ją uzgodnić z dekompozycją strukturalną.

Dotychczasowym rozważaniom można zarzucić, że nie odniosły się one *explicite* do centralnego zagadnienia związanego z wyjaśnieniami funkcjonalnymi i ich autonomią, mianowicie do kwestii *wie-*

[28] Nawiązując do rozważań Lawrence'a Shapiro (2000), można powiedzieć, że modele obliczeniowe są neutralne ze względu na te fizyczne własności, które nie są istotne czy relewantne dla wyjaśnianej zdolności poznawczej (w takim znaczeniu, w jakim na przykład posiadany kolor nie jest istotny/relewantny dla funkcjonowania otwieracza do wina). Nie są one natomiast neutralne ze względu na fizyczne własności mechanizmu, które są relewantne czy istotne dla wyjaśnianej zdolności.

lorakiej realizowalności modeli czy opisów funkcjonalnych. Modele funkcjonalne (w tym obliczeniowe) mogą być realizowane przez różniące się fizycznie mechanizmy. Tymczasem to właśnie wieloraka realizowalność miała tradycyjnie gwarantować kognitywistyce autonomię wobec neuronauki (Fodor 2008). Na pierwszy rzut oka jest jednak zupełnie niejasne, jak dokładnie zjawisko wielorakiej realizowalności miałoby uderzać w mechanistyczne ujęcie wyjaśniania w kognitywistyce. Po pierwsze, jak pokazałem, fakt, iż modele funkcjonalne są do pewnego stopnia neutralne ze względu na fizyczną budowę opisywanych przez nie mechanizmów, nie zagraża tezie, że modele funkcjonalne to w istocie szkice mechanizmów. Po drugie, poziomy mechanizmów to poziomy organizacji kompozycyjnej, a nie realizacji (por. sekcja 2.2.1; por. Craver 2007: 211–217). Między poziomami mechanizmów zachodzi relacja kompozycji, a nie bycia realizowanym.

Istnieje jednak jeden sposób, w jaki zjawisko wielorakiej realizowalności przynajmniej potencjalnie mogłoby uderzać w przyjmowane tu przeze mnie, mechanistyczne ujęcie wyjaśniania w kognitywistyce. Relacja między poziomami mechanizmów nie została co prawda wcześniej scharakteryzowana jako relacja między poziomami realizacji, lecz jako pewna forma relacji kompozycyjnej. Mimo to ktoś mógłby zaproponować, że relacja realizacji zachodzi między *zdolnościami* przypisywanymi systemom jako całościom a *mechanizmami* odpowiadającymi za te zdolności. Zdolności zaś mogą być *wielorako realizowane mechanistycznie*, w tym sensie, że (1) co najmniej dwa różne pod względem budowy i organizacji fizycznej rodzaje systemów fizycznych mogą posiadać jedną (typicznie) zdolność; (2) w każdym z tych rodzajów systemów poprawne będzie inne mechanistyczne wyjaśnienie tej zdolności. U dwóch różnych gatunków biologicznych – dwa różne typy mechanizmów neurobiologicznych mogłyby odpowiadać za tę samą zdolność poznawczą, cechującą oba te gatunki. Fakt ten miałby uderzać w twierdzenie, że zdolności są wyjaśnialne za pomocą mechanizmów neurobiologicznych, co z kolei miałoby gwarantować kognitywistyce autonomię względem neuronauki.

Także powyższa argumentacja nie zagraża jednak mechanicyzmowi i zawartemu w nim spojrzeniu na rolę neuronauki w na-

ukach kognitywnych. Można na nią odpowiedzieć na dwa sposoby. Jeden typ odpowiedzi nie jest argumentem czysto pojęciowym. Odwołuje się on raczej do pewnych metafizycznie przygodnych faktów o świecie fizycznym. Zgodnie z taką repliką możliwość i skala „występowania" wielorakiej realizowalności w świecie fizycznym jest przeceniona przez wielu filozofów umysłu. Po pierwsze, wydaje się, że autorzy opowiadający się za wieloraką realizowalnością stosują gruboziarniste kryteria indywiduacji rodzajów mentalnych, jednocześnie przyjmując nieuzasadnienie drobnoziarniste, naukowo nieakceptowalne kryteria indywiduacji rodzajów neurobiologicznych (Bechtel, Mundale 1999). Teza o wielorakiej realizowalności ma zatem u swoich podstaw pewną pomyłkę naukową, polegającą na stosowaniu zbyt szczegółowej taksonomii stanów neurobiologicznych. Po drugie, istnieją racje za tym, by twierdzić, że występują nomologiczne i „inżynieryjne" granice wielorakiej realizowalności: wiele (większość) zdolności poznawczych czy biologicznych może być realizowana przez bardzo niewielkie spektrum mechanizmów, spełniających ściśle określone ograniczenia strukturalne (Shapiro 2000; Bechtel 2008: 139–141). Obie te obserwacje prowadzą do wniosku, że, mówiąc nieco kolokwialnie, wieloraka realizowalność nie występuje „w naturze" – lub występuje jedynie w stopniu minimalnym.

Drugi typ odpowiedzi na zarzut z wielorakiej mechanistycznej realizowalności zdolności poznawczych ma z kolei naturę czysto konceptualną. Powinniśmy bowiem zapytać, dlaczego tak rozumiana wieloraka realizowalność zdolności poznawczych w ogóle miałaby stanowić problem dla mechanistycznego modelu wyjaśniania? Otóż wydaje się, że mogłoby tak być tylko w świetle ściśle określonych założeń dotyczących wzajemnych relacji różnych poziomów wyjaśniania. Argument odwołujący się do wielorakiej realizowalności będzie skuteczny tylko wtedy, gdy przyjmiemy, że wyjaśnienie zdolności przyjmuje postać redukcji interteoretycznej, rozumianej jako dedukcja jednej teorii z drugiej (Fodor 2008; por. Bechtel 2008: 155–157). Taka redukcja wymaga wszakże sformułowania praw czy definicji mostowych, specyfikujących definicyjne związki terminów teorii redukowanej i redukującej. Wieloraka realizowalność wyklucza możliwość stworzenia takich interteoretycznych pojęciowych

„mostów": egzemplifikacja żadnej własności neurobiologicznej nie jest zarazem konieczna i wystarczająca dla zaistnienia określonej własności (zdolności) psychologicznej czy poznawczej (por. Fodor 2008). Jeśli zatem przyjmiemy tego rodzaju model redukcji interteoretycznej, to możliwość wyjaśnienia zdolności poznawczych przez odwołanie do ich mechanizmów będzie problematyczna.

Trzeba jednak zauważyć, że mechanicyzm wcale nie przyjmuje tego rodzaju teorii redukcji (Craver 2007: 233–246; Bechtel 2008: 142–157). Nie ma w nim mowy o tym, aby predykaty opisujące zdolności systemów jako całości (poziom wyższy) miały być definiowalne za pomocą predykatów opisujących składowe i organizację mechanizmu (poziom niższy). Mechanistyczny model wyjaśniania nic takiego nie zakłada. Z punktu widzenia mechanicyzmu nic nie stoi na przeszkodzie, aby stwierdzić, że jedna zdolność ma dwa różne wyjaśnienia mechanistyczne, a poprawność każdego z nich jest zrelatywizowana do, dla przykładu, danego gatunku biologicznego. W fakcie tym nie ma nic problematycznego czy „tajemniczego". Tym samym zarzut z wielorakiej mechanistycznej realizowalności zdolności mentalnych jest nietrafiony i nie można na jego podstawie argumentować za autonomią kognitywistyki względem neuronauki.

Obrona mechanicyzmu przed zarzutami odwołującymi się do funkcjonalnej natury wyjaśniania w kognitywistyce oraz wielorakiej realizowalności zdolności poznawczych zamyka przedstawioną tu rekonstrukcję mechanistycznego modelu wyjaśniania. W kolejnych rozdziałach tej książki będę przyjmować, że zasadniczym czy fundamentalnym rodzajem wyjaśniania w kognitywistyce jest wyjaśnianie za pomocą mechanizmów. Ujmując to w kategoriach bardziej normatywnych, duża część praktyki epistemicznej kognitywistów, która *zasługuje* na miano praktyki pełnoprawnie eksplanacyjnej, opiera się na strukturalnej i funkcjonalnej dekompozycji systemu poznawczego, wykonywanej w celu odkrycia mechanizmów stojących u podstaw zdolności poznawczych. Następny rozdział poświęcę zastosowaniu tej ogólnej idei do kwestii eksplanacyjnej wartości reprezentacji mentalnych. Spróbuję w nim wykorzystać mechanistyczny model wyjaśniania w celu sformułowania kryterium bycia wyjaśnieniem reprezentacyjnym.

ROZDZIAŁ 3

Reprezentacje i mechanicyzm. Problem mechanizmów reprezentacyjnych

3.1. Wyjaśnianie mechanistyczne za pomocą reprezentacji: uwagi wstępne

3.1.1. Wyjaśnianie reprezentacyjne jako wyjaśnianie za pomocą mechanizmów reprezentacyjnych

Dysponując określoną koncepcją wyjaśniania w kognitywistyce, mogę powrócić do zagadnienia statusu eksplanacyjnego reprezentacji mentalnych w naukach kognitywnych. Na przestrzeni tego oraz następnych rozdziałów będę przyjmować założenie, że postulując „wewnętrzne", zlokalizowane w mózgu reprezentacje, naukowcy są zaangażowani w odkrywanie mechanizmów poznania. Reprezentacje mentalne grają rolę w ramach wewnętrznej, mechanistycznej architektury systemu poznawczego. Mówiąc inaczej, proponuję uznać, że formułowane w kognitywistyce, odwołujące się do pojęcia reprezentacji wyjaśnienia różnych zjawisk poznawczych powinny być interpretowane jako pewne rodzaje wyjaśnień mechanistycznych. Ujmując tę myśl precyzyjniej, proponuję uznać, że z punktu widzenia mechanicyzmu wyjaśnienia reprezentacyjne odwołują się do specyficznego rodzaju mechanizmów poznawczych, których działanie opiera się – jakkolwiek by to twierdzenie rozumieć – na wewnętrznych reprezentacjach. Mechanizmy tego rodzaju będę tu po prostu nazywał „mechanizmami reprezentacyjnymi". Czy istnieją zjawiska poznawcze (zdolności w sensie Roberta Cumminsa), które u swoich podstaw mają działanie mechanizmów reprezentacyjnych? Czy istnieją w kognitywistyce eksplananda domagające się wyjaśniania za pomocą mechanizmów reprezentacyjnych? Jakie to eksplananda?

Zauważmy, że powyższe pytania koncentrują się na problemie innym niż ten, który wyznacza główny przedmiot zainteresowania tej książki. Dotyczą one bowiem problemu statusu eksplanacyjnego reprezentacji w jego *przedmiotowym* wymiarze, to znaczy zagadnienia, czy i w jakim zakresie poprawne wyjaśnienia zjawisk umysłowych mają charakter reprezentacyjny. Tymczasem właściwym przedmiotem prowadzonych przeze mnie rozważań jest *metaprzedmiotowe* pytanie o to, czym w ogóle są – czy też na czym polegają – wyjaśnienia reprezentacyjne. Z przyjmowanej tu perspektywy wartość mechanicyzmu polega na tym, że pozwala on dookreślić właśnie ten metaprzedmiotowy problem. Pozwala mianowicie zamienić ogólne pytanie: „Czym są wyjaśnienia reprezentacyjne?", na bardziej konkretne i precyzyjne: „Czym są *mechanistyczne* wyjaśnienia reprezentacyjne?". Sądzę, że udzielenie odpowiedzi na to ostatnie pytanie stanowi klucz do rozwiązania problemu reprezentacji w jego metaprzedmiotowym wymiarze.

Biorąc pod uwagę te wstępne uwagi, przyjmijmy następujące, bardzo ogólne rozumienie mechanistycznego wyjaśniania reprezentacyjnego:

> Mechanistyczne wyjaśnienie *W* eksplanandum *E* jest reprezentacyjne wtedy i tylko wtedy, gdy *W* wyjaśnia *E* za pomocą mechanizmu reprezentacyjnego.

Tym, co stanowi o reprezentacyjnej naturze wyjaśnienia, jest więc reprezentacyjna natura mechanizmu, na który powołuje się ono jako eksplanans. Fakt ten pozwala raz jeszcze dookreślić pytanie o naturę wyjaśnień reprezentacyjnych, tym razem przedstawiając je jako pytanie: „Czym są mechanizmy reprezentacyjne?". Mechanizmy te charakteryzują się tym, że ich działanie „opiera się" na wewnętrznych reprezentacjach. Na czym jednak polega „opieranie się" działania mechanizmu na reprezentacjach? Czym mechanizmy reprezentacyjne różnią się od tych, którym taki status nie przysługuje?

Mówiąc bardzo szeroko, możemy przyjąć, że jeśli pewien byt postulowany ma mieć wartość eksplanacyjną w ramach wyjaśnienia mechanistycznego – czyli ma stanowić element tego wyjaśnienia –

to będzie on ją posiadał jako składowa stosownego mechanizmu. To z kolei znaczy, że byt ten będzie dysponował wartością eksplanacyjną jako *komponent* lub *operacja* wykonywana przez komponent mechanizmu. Ta ogólna idea powinna przyświecać myśleniu o reprezentacjach, o ile te są rozumiane właśnie jako byty postulowane w ramach wyjaśnień mechanistycznych. Możemy w takim duchu przyjąć, że „opieranie się" przez mechanizm na reprezentacjach polega na fakcie, iż komponent (lub grupa komponentów) tego mechanizmu wykonuje w jego ramach *operację polegającą na reprezentowaniu czegoś*. Inaczej mówiąc, mechanizm reprezentacyjny to taki mechanizm, w którym jeden lub więcej komponentów jest zaangażowanych w realizowanie funkcji polegającej na reprezentowaniu czegoś. Tak jak działanie mechanizmu odpowiedzialnego za transport składników odżywczych w organizmie opiera się na komponencie wykonującym operację pompowania, tak mechanizmy reprezentacyjne będą wykorzystywać komponent(y) zaangażowany(e) w reprezentowanie czegoś. Reprezentowanie – w eksplanacyjnie wartościowym sensie – stanowi zatem operację wykonywaną przez pewien komponent mechanizmu. Komponent wykonujący taką operację jest reprezentacją, czy też, bardziej precyzyjnie, *nośnikiem* reprezentacji[1].

Warto podsumować dotychczasowe ustalenia. Kluczem do zrozumienia natury wyjaśnienia reprezentacyjnego jest zrozumienie natury mechanistycznego wyjaśnienia reprezentacyjnego. Mechanistyczne wyjaśnienie reprezentacyjne to takie, które wyjaśnia dane zjawisko za pomocą mechanizmu reprezentacyjnego. Ten ostatni zaś to taki mechanizm, którego działanie opiera się na wewnętrznej reprezentacji. To z kolei oznacza, że co najmniej jeden komponent ta-

[1] Zauważmy, że twierdzenia te pozostają neutralne w kwestii tego, co może podlegać reprezentowaniu w mechanizmie. Reprezentowane może być zarówno środowisko zewnętrzne względem systemu poznawczego, jak i ciało własne (jego morfologia, motoryka czy stany, w jakich znajdują się narządy wewnętrzne). Przedstawione tu rozumienie mechanizmów reprezentacyjnych nie przesądza tej kwestii.

kiego mechanizmu wykonuje operację polegającą na reprezentowaniu czegoś. Bardziej technicznie[2]:

> Mechanistyczne wyjaśnienie *W* eksplanandum *E* jest reprezentacyjne wtedy i tylko wtedy, gdy (1) *W* wyjaśnia *E* przez odwołanie do mechanizmu *M*, na który składają się zorganizowane komponenty {*X1*, *X2*,..., *Xm*} i działania tych komponentów {*φ1*, *φ2*, ..., *φn*} oraz (2) *E* przypisuje co najmniej jednemu z komponentów {*X1*, *X2*, ..., *Xm*} mechanizmu *M* operację polegającą na reprezentowaniu[3].

Można zatem wyciągnąć wniosek, że problem eksplanacyjnego statusu reprezentacji w jego metaprzedmiotowym wymiarze sprowadza się do zagadnienia *reprezentowania jako operacji wykonywanej przez komponent mechanizmu*. Na tym jednak kończy się ta łatwa część aplikowania koncepcji wyjaśniania mechanistycznego do kwestii reprezentacjonizmu. W poprzednim rozdziale przyjąłem, że operacja komponentu mechanizmu to sposób, w jaki komponent ten przyczynia się do ogólnego funkcjonowania całości. Tym samym operacja komponentu to jego funkcja w ramach mechanizmu. Na czym polega jednak funkcjonowanie jako reprezentacja (pełnienie funkcji reprezentacji) w ramach mechanizmu poznawczego? Jaki musi być profil funkcjonalny komponentu mechanizmu, abyśmy mogli w uzasadniony sposób uznać, że komponent ten jest zaangażowany w realizowanie operacji reprezentowania czegoś (jest nośnikiem reprezentacji)? Jeśli przyjmiemy perspektywę mechanistyczną, to okazuje się, że właśnie te pytania wyznaczają sedno metaprzedmiotowego problemu reprezentacji mentalnych. Tym, czego tu potrzeba, jest koncepcja służenia czy funkcjonowania jako reprezentacja w ramach mechanizmu poznawczego. Potrzebujemy zatem *funkcjonalnej* koncepcji reprezentacji, rozumianej jako koncepcja specyfi-

[2] Powyższy sposób symbolicznego oznaczenia komponentów i ich działań został zaczerpnięty z prac Cravera (por. np. Craver 2007: 6–7).

[3] Tym samym wśród operacji {*φ1*, *φ2*, ... , *φn*} znajduje się także opracja reprezentowania.

kująca funkcjonalne kryteria bycia reprezentacją w mechanizmie poznawczym.

Warto jeszcze zawrócić uwagę na fakt, że termin „mechanizm reprezentacyjny" da się interpretować w dwojaki sposób, co może generować ekwiwokacje. Dobrze będzie *explicite* wyróżnić te dwa znaczenia i zaznaczyć wyraźnie, którym z nich będę się posługiwać w dalszej części pracy. W pierwszym znaczeniu, mechanizmy reprezentacyjne to takie, które odgrywają swoją rolę eksplanacyjną (między innymi) dzięki temu, że reprezentacje wchodzą w ich skład. To znaczy reprezentacje (nośniki reprezentacji) to komponenty takich mechanizmów. Mechanizm reprezentacyjny w takim znaczeniu ma status *eksplanansu*. Wyjaśnia on pewne zjawisko między innymi dzięki temu, że „korzysta" z komponentu odgrywającego w nim rolę reprezentacji. Pamiętajmy jednak o hierarchicznej, kompozycyjnej naturze mechanizmów. Operacje wykonywane przez komponenty mechanizmu mogą stać się *eksplanandami* wyjaśnianymi na niższym poziomie. Właśnie tu pojawia się drugie potencjalne znaczenie terminu „mechanizm reprezentacyjny". Otóż reprezentowanie czegoś przez komponent mechanizmu wyższego poziomu może samo być wyjaśnione mechanistycznie. Jeśli pewien komponent wykonuje operację reprezentowania czegoś, to jego zdolność do reprezentowania może być wyjaśniona przez wskazanie mechanizmu na niższym poziomie organizacji. Także ten mechanizm *wyjaśniający reprezentowanie* (a nie: korzystający z reprezentacji) może być potencjalnie zakwalifikowany jako „reprezentacyjny". Mając na uwadze to rozróżnienie, trzeba wyraźnie zaznaczyć, że w książce tej wykorzystuję termin „mechanizm reprezentacyjny" w pierwszym wymienionym znaczeniu. W przyjętym tu przeze mnie rozumieniu, mechanizmy reprezentacyjne to nie mechanizmy *wyjaśniające* zdolność do reprezentowania, lecz *wykorzystujące reprezentacje* i wyjaśniające w ten sposób pewne inne zjawisko (zjawiskiem takim może być *prima facie* dowolna zdolność należąca do klasy eksplanandów nauk kognitywnych). Interesuje mnie tu reprezentowanie jako eksplanans (element eksplanansu), a nie eksplanandum wyjaśnienia mechanistycznego.

3.1.2. Mechanicyzm i problem reprezentacji: kilka kwestii dodatkowych

Zanim przejdę do poszukiwania koncepcji mechanizmów reprezentacyjnych, chcę poczynić kilka – jak sądzę, istotnych – uwag dodatkowych. Wydaje się, że mechanistyczna perspektywa pozwala zmodyfikować, postawić w nowym świetle lub zaproponować nowe rozwiązania przynajmniej niektórych zaznaczonych w rozdziale 1 szczegółowych zagadnień związanych z reprezentacjami mentalnymi i rolą, jaką mają one do spełnienia w kognitywistyce. Uważam, że kwestie są na tyle znaczące, iż warto je wymienić i omówić. Poniższa lista daje wyobrażenie, w jaki sposób mechanistyczny model wyjaśniania może wpłynąć na krajobraz współczesnych debat na temat reprezentacji wewnętrznych (por. także Gładziejewski 2013a).

a) Problem roli funkcjonalnej reprezentacji a problem treści intencjonalnej

Jak już zostało wspomniane w rozdziale 1, w ciągu ostatnich paru dekad wielu naturalistycznie zorientowanych filozofów umysłu było zaprzątniętych problemem naturalizacji intencjonalności. W takim klimacie intelektualnym „problem reprezentacji" na ogół rozumiano jako problem dostarczenia naturalistycznej koncepcji *treści intencjonalnej*. Do centralnych punktów spornych w dyskusji nad naturą treści należała chociażby opozycja między teoriami internalistycznymi a eksternalistycznymi. Filozoficzne próby rozwiązania tego sporu często opierały się na wyrafinowanych eksperymentach myślowych, których znaczenie dla realnej praktyki badawczej i eksplanacyjnej kognitywistów okazywało się raczej marginalne. Co więcej, wyniki tego rodzaju rozważań na temat treści reprezentacji na ogół nie przekładały się też na kwestię tego, co to znaczy odgrywać *rolę funkcjonalną* reprezentacji w ramach mechanizmu poznawczego. Nie dawały one żadnych wskazówek, jeśli chodzi o kwestię funkcjonalnych kryteriów bycia reprezentacją. Tymczasem z mechanistycznego punktu widzenia to właśnie to ostatnie zagadnienie okazuje się kluczowe, kiedy pytamy o eksplanacyjny status reprezentacji w kognitywistyce. Wydaje się zatem, że problem tego, co nadaje reprezen-

tacjom treść, jest przynajmniej częściowo odrębny i niezależny od problemu użyteczności eksplanacyjnej reprezentacji w naukach kognitywnych. Jeśli interesuje nas nie tyle naturalizacja treści intencjonalnej, co raczej zagadnienie wartości eksplanacyjnej reprezentacji dla nauk kognitywnych, to powinniśmy przenieść ciężar rozważań z problemu treści na problem roli funkcjonalnej reprezentacji. Mówiąc inaczej, powinna nas interesować nie tyle treść reprezentacji, co raczej rola funkcjonalna pełniona w systemie poznawczym przez nośnik tej treści: przez to, co reprezentuje.

Rzecz jasna oba te zagadnienia – problem treści i problem funkcji reprezentacji – nie powinny być traktowane jako *całkowicie* odrębne i niezależne. Mówiąc na przykład, że pewien stan wewnętrzny pełni funkcję reprezentacji, można mieć między innymi na myśli to, iż rola ta jest (współ)determinowana przez treść tego stanu: funkcjonować jako reprezentacja to przecież pełnić takie, a nie inne role dlatego, że posiada się określoną treść (por. Ramsey 2007: 18–20). Jeśli tak, to mechanicysta powinien dysponować przynajmniej wstępną czy szkicową (a także, rzecz jasna, naturalistyczną) koncepcją treści[4]. Prędzej czy później trzeba będzie bowiem stanąć przed problemem tego, czym jest treść reprezentacji zawartych w mechanizmach reprezentacyjnych oraz jak ta treść jest determinowana. Tym samym przyjęcie perspektywy mechanistycznej w żadnym wypadku nie oznacza wcale, że problem treści staje się mało istotny lub możliwy do całkowitego pominięcia w myśleniu o reprezentacjach mentalnych[5]. Jednakże, mimo tych zastrzeżeń, należy wyraźnie zazna-

[4] Kwestia ta komplikuje się dodatkowo, jeśli przypomnimy sobie, że projekt naturalizacji intencjonalności na ogół był rozumiany (bardziej lub mniej otwarcie) jako projekt naturalizacji treści postaw propozycjonalnych. Te ostatnie to stany, w jakich znajdują się *osoby*. Niewykluczone jednak, że treści przypisywane stanom, w których znajdują się (subosobowe) komponenty mechanizmów poznawczych będą zasadniczo różniły się w stosunku do treści, które są przypisywane postawom propozycjonalnym. Jakie konsekwencje miałaby taka sytuacja dla projektu naturalizacji intencjonalności? Temu oraz pokrewnym zagadnieniom jest w całości poświęcony rozdział 5.

[5] Należy poczynić tu dwa dodatkowe zastrzeżenia. Po pierwsze, niektóre koncepcje powstałe w kontekście problemu treści mogą być potraktowane jako teorie dające pewnego rodzaju odpowiedź na pytanie o to, co to znaczy funkcjonować

czyć, że w centrum dalszych rozważań przedstawionych w tej książce znajduje się kwestia funkcjonowania czy służenia jako reprezentacja w ramach mechanizmu poznawczego. Problem tego, co nadaje reprezentacjom treść, zostanie tu podjęty jedynie szkicowo, a skupię się na roli pełnionej w systemie (czy mechanizmie) przez nośniki. To ten ostatni problem stanowi bowiem klucz do rozwiązania problemu natury wyjaśnień reprezentacyjnych.

b) Wyjaśnianie mechanistyczne a reprezentacjonizm globalny i lokalny
W rozdziale 1 wprowadziłem rozróżnienie na globalną oraz lokalną wersję reprezentacjonizmu i stanowiska opozycyjnego. Reprezentacjonizm oraz antyreprezentacjonizm mogą być stanowiskami dotyczącymi działania systemu poznawczego jako takiego, bądź stanowiskami mającymi jedynie lokalne zastosowanie, zrelatywizowane do określonego eksplanandum czy klasy eksplanandów. Pierwsza, globalna wersja (anty)reprezentacjonizmu niesie ze sobą niebezpieczeństwo związane z nieuprawnioną generalizacją własnego stanowiska. Jest przecież empirycznie możliwe, że część zdolności umysłowych powinna być wyjaśniana reprezentacyjnie, a pozostała część bez po-

jako reprezentacja w ramach mechanizmu. Ramsey (2007: 127–131) na przykład argumentuje, że tak właśnie można traktować teorie, które współzmiennościową koncepcję treści uzupełniają elementami o charakterze teleologicznym (por. sekcja 3.3.1). Nie można więc *a priori* założyć, że powstałe do tej pory naturalistyczne koncepcje treści intencjonalnej są kompletnie nieważne dla problemu funkcjonalnych kryteriów bycia reprezentacją. Po drugie, odróżnienie problemu treści od problemu funkcjonowania jako reprezentacja nie implikuje w żadnym wypadku, że powinniśmy przyjąć „wąską" koncepcję treści, zgodnie z którą treść reprezentacji (nośnika reprezentacji) jest konstytuowana przez jej role funkcjonalne. Chodzi raczej o to, że obydwa zagadnienia mogą być traktowane oddzielnie. Propozycje dotyczące tego, jakie własności nadają czemuś treść intencjonalną, na ogół nie będą przesądzały o tym, w jaki sposób stany posiadające tę treść funkcjonują w ramach określonego mechanizmu (a nawet o tym, czy ich role funkcjonalne w ogóle polegają na reprezentowaniu w jakimś wartościowym eksplanacyjnie sensie). Jednocześnie nie można wykluczyć, że funkcjonalna koncepcja reprezentacji – koncepcja dotycząca tego, na czym polega pełnienie funkcji reprezentacji w mechanizmie poznawczym – będzie pozostawać przynajmniej częściowo neutralna, jeśli chodzi o problem tego, co nadaje reprezentacjom treść.

woływania się na reprezentacje. Przyjęcie mechanistycznej perspektywy ma tę zaletę, że dość naturalnie sprzyja postrzeganiu sporu reprezentacjonizm–antyreprezentacjonizm jako szeregu szczegółowych (lokalnych) sporów dotyczących za każdym razem konkretnego zjawiska poznawczego czy klasy takich zjawisk. Fakt ten wynika ze sposobu, w jaki w koncepcji wyjaśnienia mechanistycznego rozumie się same mechanizmy. Mechanizmy są wszakże po części indywiduowane przez odniesienie do zjawisk, które mają one wyjaśniać. Za każdym razem mówimy o mechanizmie *czegoś*, gdzie „coś" jest wyznaczane przez eksplanandum, które podlega wyjaśnieniu. Różne zjawiska są na ogół wyjaśniane za pomocą różnych mechanizmów. Ta ogólna idea stosuje się oczywiście także do kognitywistyki. Konkretne eksplananda nauk kognitywnych będą wyjaśniane za pomocą różnych, odrębnych mechanizmów. Bardzo prawdopodobne, że tylko część spośród tych eksplanandów może zostać wyjaśniona przez odwołanie do mechanizmów reprezentacyjnych.

Powyższe twierdzenie naturalnie rodzi pytanie: które zjawiska domagają się wyjaśnienia w terminach reprezentacyjnych (za pomocą mechanizmów reprezentacyjnych), a które nie? Istotnym filozoficznym głosem w tej sprawie był wspomniany już w rozdziale 1, opublikowany w latach dziewięćdziesiątych poprzedniego wieku artykuł Andy'ego Clarka i Josefy Toribio (1994). Przypomnijmy, że autorzy ci zaproponowali szereg kryteriów służących odróżnianiu zdolności poznawczych, które są „złaknione reprezentacji" (*representation-hungry*), od takich, które z reprezentacji korzystać nie muszą. Te pierwsze są związane z koniecznością behawioralnego lub poznawczego orientowania się przez organizm względem abstrakcyjnych lub nieobecnych (czasoprzestrzennie oddalonych) obiektów i własności. Inna (choć, jak się wydaje, pokrewna) propozycja mogłaby głosić, że „wyższe" zdolności poznawcze mają u swoich podstaw mechanizmy reprezentacyjne, natomiast mechanizmy wyjaśniające zdolności „niższe" (jakkolwiek by taką opozycję „niższe–wyższe" rozumieć) nie są – czy nie muszą być – reprezentacyjne. Problemu, które zjawiska (jeśli jakiekolwiek) powinny być wyjaśniane za pomocą mechanizmów reprezentacyjnych, nie sposób jednak rozstrzygnąć konkluzywnie na gruncie czysto konceptualnych

rozważań. To, co okazuje się tu istotne, to jedynie fakt, że patrząc przez pryzmat mechanicyzmu, spór reprezentacjonizm–antyreprezentacjonizm powinien być rozstrzygany w sposób zrelatywizowany do konkretnych eksplanandów lub klas eksplanandów. Nie jest się (anty)reprezentacjonistą po prostu, lecz raczej (anty)reprezentacjonistą w stosunku do konkretnej zdolności poznawczej czy klasy takich zdolności[6].

c) Mechanicyzm a realistyczne podejście do reprezentacji

Wspomniany już w rozdziale 1 Anthony Chemero (2009: 67–83) odróżnia reprezentacjonizm jako stanowisko „epistemologiczne" od reprezentacjonizmu pojętego jako stanowisko metafizyczne. Takie samo odróżnienie stosuje się do antyreprezentacjonizmu. W pierwszej, epistemologicznej wersji, oba stanowiska – reprezentacjonizm i antyreprezentacjonizm – dotyczą sposobu wyjaśniania systemu poznawczego, który jest w jakimś sensie najlepszy czy optymalny ze względu na potrzeby i możliwości poznawcze istot ludzkich. W drugiej, metafizycznej wersji, stanowiska te dotyczą rzeczywistej natury systemu poznawczego, czyli tego, czy system ten ma naturę reprezentacyjną, czy nie. Chemero przyjmuje, że epistemologiczne i metafizyczne sposoby rozumienia (anty)reprezentacjonizmu należy uznać za niezależne. Według niego zwolennik lub przeciwnik reprezentacjonizmu może bronić czysto epistemologicznego stanowiska, pozostając agnostykiem w metafizycznych kwestiach dotyczących natury (czy sposobu działania) systemu poznawczego. Można bronić epistemologicznego reprezentacjonizmu, dopuszczając jednocześ-

[6] Oczywiście mechanicyzm nie wyklucza możliwości, że ostatecznie prawdziwa okaże się globalna wersja reprezentacjonizmu lub antyreprezentacjonizmu. Być może istnieje pewien rodzaj mechanizmów reprezentacyjnych, które odpowiadają za ogromną część zjawisk poznawczych – na tyle sporą, by można było uznać system poznawczy za „globalnie" reprezentacyjny. Sytuacja może okazać się jednak dokładnie przeciwna i prawdziwa będzie globalna forma antyreprezentacjonizmu. Mechanicyzm nie wyklucza żadnej z tych możliwości. Zaleta perspektywy mechanistycznej nie polega na tym, że wyklucza ona globalną wersję (anty)reprezentacjonizmu, ale na tym, że dzięki jej przyjęciu naturalnie unika się sytuacji, w której o prawdziwości takiej globalnej wersji jednego z tych stanowisk przesądza się *a priori*.

nie możliwość, że rola eksplanacyjna reprezentacji pozostaje czysto instrumentalna, to znaczy, że reprezentacje stanowią jedynie epistemicznie użyteczne fikcje.

Mechanicyzm jest nie do pogodzenia z tak silnym odróżnieniem kwestii epistemologicznych i metafizycznych. Co prawda wśród samych mechanicystów istnieje rozłam między zwolennikami „ontycznego" a zwolennikami „epistemicznego" ujęcia wyjaśniania mechanistycznego. Mówiąc w uproszczeniu, zwolennicy podejścia ontycznego (por. np.: Craver 2007: 107–162; Machamer, Darden, Craver 2011) uważają, że zjawiska wyjaśniają *same mechanizmy* jako byty w świecie; natomiast zwolennicy podejścia epistemicznego (por. np.: Bechtel 2008: 17–22; Wright 2012) uznają, że wszelkie wyjaśnienia są epistemicznymi produktami nauki, a zatem wyjaśnienia mechanistyczne to nie same mechanizmy, lecz modele lub innego rodzaju *naukowe reprezentacje* mechanizmów[7]. Jak się jednak wydaje, obie strony tego sporu akceptują jedno zasadnicze założenie (por. rozdział 2, przypis 15). Nawet jeśli przyjmiemy epistemiczną koncepcję wyjaśniania, to i tak musimy założyć, że jednym z kryteriów czy norm poprawności wyjaśnienia mechanistycznego jest to, czy wyjaśnienie to trafnie oddaje *naturę rzeczywistego mechanizmu* stojącego u podstaw danego zjawiska (Bechtel 2007; por. także Illari 2013). Wyjaśnienie będzie tym lepsze, im bardziej adekwatnie odzwierciedla stosowny mechanizm: jego komponenty, ich działania oraz organizację[8]. Twierdzenie to stosuje się rzecz jasna także do wyjaśnień w kognitywistyce. Nawet potraktowane jako twory epistemiczne, wyjaśnienia mechanistyczne w naukach kognitywnych są

[7] Precyzyjniej, w ujęciu epistemicznym wyjaśnienia mogą być rozumiane jako inferencyjne operacje na naukowych reprezentacjach mechanizmów, które to operacje dają nam zrozumienie, jak reprezentowany mechanizm umożliwia zachodzenie zjawiska stanowiącego eksplanandum (Wright 2012).

[8] Wyobraźmy sobie, że dysponujemy dwoma alternatywnymi wyjaśnieniami mechanistycznymi, które pozwalają równie skutecznie przewidywać zachowanie pewnego systemu fizycznego. Które z tych alternatywnych wyjaśnień mamy uznać za poprawne lub przynajmniej bliższe poprawności? Dość oczywistym kryterium różnicującym moc eksplanacyjną jest to, które z proponowanych wyjaśnień bardziej adekwatnie oddaje rzeczywistą organizację mechanizmu stojącego u podstaw wyjaśnianego zjawiska.

poprawne o tyle, o ile oddają funkcjonalną, strukturalną i organizacyjną „architekturę" rzeczywistych mechanizmów poznawczych. Warunkiem poprawności wyjaśnienia pewnego zjawiska za pomocą mechanizmu reprezentacyjnego jest to, by działanie mechanizmu odpowiadającego za to zjawisko rzeczywiście opierało się na wewnętrznych reprezentacjach. Twierdzenia epistemologiczne dotyczące tego, jakie wyjaśnienia systemu poznawczego są poprawne czy pożądane, pociągają za sobą twierdzenia metafizyczne dotyczące tego, jak dany mechanizm rzeczywiście działa. Jeśli za dane zjawisko odpowiada mechanizm reprezentacyjny, nie można *poprawnie* wyjaśnić tego zjawiska za pomocą mechanizmu nieposługującego się reprezentacjami. Postulowana przez Chemero niezależność epistemologicznego i metafizycznego sformułowania (anty)reprezentacjonizmu będzie zatem nie do utrzymania dla mechanicysty. Z tego też powodu mechanicyzm naturalnie sprzyja *realistycznemu* podejściu do reprezentacji i jest niekompatybilny z instrumentalizmem. Marcin Miłkowski (2013: vii), podejmując w kontekście mechanicyzmu zagadnienie wyjaśnień obliczeniowych w kognitywistyce, stwierdza, że umysł (system poznawczy) może być wyjaśniony obliczeniowo, ponieważ jest obliczeniowy. Analogicznie z reprezentacjami. System poznawczy może być wyjaśniony reprezentacyjnie tylko wtedy, gdy *jest* on systemem reprezentacyjnym.

d) Wyjaśnienia dynamiczne a reprezentacjonizm

W rozdziale 1 stwierdziłem, że na podstawie literatury przedmiotu można uznać, iż teoria systemów dynamicznych to sprzymierzeniec współczesnego antyreprezentacjonizmu. Ten ostatni bywa uznawany za naturalny rezultat przyjęcia dynamicznego podejścia do wyjaśniania zdolności poznawczych. Rozumienie systemu poznawczego jako sprzężonego ze środowiskiem systemu dynamicznego ma być alternatywą dla postrzegania go jako reprezentacyjnego (van Gelder 1995; Chemero 2009, 2014). Jeśli spojrzymy na te twierdzenia z perspektywy mechanistycznego modelu wyjaśniania, okazuje się, że związek między podejściem dynamicznym a antyreprezentacjonizmem nie jest tak oczywisty. Istnieje bowiem grupa zwolenników mechanistycznego podejścia do wyjaśniania, która przekonują-

co broni twierdzenia, że wyjaśnienia korzystające z teorii systemów dynamicznych mogą, a nawet powinny zostać uznane za formę wyjaśnień mechanistycznych (Bechtel, Abrahamsen 2010; Kaplan, Bechtel 2011; Zednik 2011). Zgodnie z ich linią argumentacji jeśli modele dynamiczne mają w ogóle dostarczać *wyjaśnień*, to muszą to być wyjaśnienia mechanistyczne. Jeśli bowiem przyjmiemy, że modele dynamiczne nie niosą ze sobą zobowiązań dotyczących wewnętrznej, mechanistycznej organizacji wyjaśnianego systemu, to zawsze mogą okazać się one predykcyjnie użytecznymi fikcjami. Jednak predykcja nie jest ani koniecznym, ani wystarczającym warunkiem wyjaśniania. Powinniśmy raczej uznać, że modele dynamiczne wyjaśniają zjawiska tylko o tyle, o ile reprezentują one rzeczywiste przyczynowe podstawy zachowania modelowanego systemu. A to będzie możliwe wtedy, gdy potraktujemy je jako opisy (modele) stosownych mechanizmów. Z takiego punktu widzenia równania zawarte w wyjaśnieniach dynamicznych opisują (powinny być interpretowane jako opisujące) wewnętrzną dynamikę mechanizmów. Dokładniej, równania te specyfikują wzorce interakcji zachodzących między komponentami mechanizmów.

Co to wszystko mówi o relacji między antyreprezentacjonizmem a dynamicznym podejściem do wyjaśniania w kognitywistyce? Załóżmy, że formułujemy koncepcję pokazującą, pod jakimi warunkami wyjaśnienie mechanistyczne jest reprezentacyjne. Jednocześnie przyjmijmy, że wyjaśnienia mechanistyczne mogą być formułowane w matematycznym języku dostarczanym przez teorię systemów dynamicznych. Jeśli tak, to *prima facie* powinna zachodzić możliwość opisywania działania mechanizmów reprezentacyjnych za pomocą kategorii dynamicznych. Tym samym, z punktu widzenia mechanicyzmu, bycie zwolennikiem perspektywy dynamicznej w wyjaśnianiu zjawisk poznawczych nie przesądza jeszcze o byciu antyreprezentacjonistą[9]. Wydaje się, że reprezentacjonizm i podejście dynamiczne są możliwe do pogodzenia.

[9] Można tę myśl ująć nieco inaczej. Nie ulega wątpliwości, że podejście dynamiczne daje nam pewne pojęcie o tym, jak mogłyby wyglądać nieodwołujące się do reprezentacji wyjaśnienia kognitywistyczne. Chodzi jednak o to, że fakt, iż pewne wyjaśnienie zostało sformułowane w języku teorii systemów

3.2. Funkcjonalne rozumienie reprezentacji i metoda Ramseya

3.2.1. Reprezentacje jako funkcjonalne komponenty mechanizmów. W poszukiwaniu funkcjonalnej koncepcji reprezentacji

Wróćmy teraz do głównego wątku rozważań, czyli pytania o naturę wyjaśnień reprezentacyjnych. Aplikacja mechanicyzmu do problemu eksplanacyjnego statusu reprezentacji mentalnych doprowadziła mnie do sformułowania wniosku, że aby rozwiązać tę ostatnią kwestię, potrzeba funkcjonalnej koncepcji reprezentacji. Należy odpowiedzieć na pytanie o to, co to znaczy pełnić funkcję reprezentacji w ramach mechanizmu poznawczego. Reprezentacje wyjaśniają zjawiska o tyle, o ile *służą jako* reprezentacje w ramach mechanizmu (czyni on z nich „użytek" jako z reprezentacji). Pytanie o ich eksplanacyjną rolę w naukach kognitywnych w istocie dotyczy więc tego, co to znaczy, że komponent mechanizmu wykonuje operację polegającą na reprezentowaniu czegoś.

Trzeba zauważyć, że taka motywowana mechanicyzmem perspektywa nie jest całkowicie odosobniona we współczesnej literaturze dotyczącej reprezentacji mentalnych. Nie jest bowiem tak, że nikt do tej pory nie dostrzegł znaczenia problemu funkcji pełnionych przez reprezentacje w systemie poznawczym.

Jak już wspomniałem w rozdziale 1, w trakcie ostatnich dekad na styku filozofii i kognitywistyki rosnącą popularność zaczęły zyskiwać stanowiska opierające się na ogólnej obserwacji, że system poznawczy to ucieleśniony, działający byt, zakorzeniony w określonym środowisku. Według zwolenników takiego podejścia większość „klasycznych" teorii powstałych w ramach kognitywistyki całkowicie pomijała lub jedynie częściowo, niewystarczająco doceniała wagę tego faktu. Teorie tego rodzaju nie brały (w wystarczającym stopniu) pod uwagę, że procesy poznawcze dokonują się w ucieleśnionym systemie, a ich pierwotną czy fundamentalną funkcją jest sprawne, dy-

dynamicznych, nie przesądza jeszcze o tym, że jest to wyjaśnienie, które nie postuluje reprezentacji.

namiczne planowanie i kontrolowanie działań tego systemu w określonym środowisku. Remedium na taki stan rzeczy mają stanowić alternatywne koncepcje, budowane w „ucieleśnionym", „enaktywnym", „ugruntowanym" czy „zakorzenionym" duchu. Co istotne, teorie tego rodzaju często bywają łączone z (preskryptywnym) *antyreprezentacjonizmem*. Zwolennicy perspektywy skoncentrowanej na ucieleśnieniu, działaniu czy roli środowiska w poznaniu bardzo często odrzucają także pojęcie reprezentacji jako ważne narzędzie eksplanacyjne nauk kognitywnych. Przy bliższym spojrzeniu stosunek takich nowszych podejść w kognitywistyce do reprezentacji mentalnych okazuje się jednak bardziej złożony. Z jednej strony stanowiska „rewizjonistyczne" rzeczywiście postulują całkowite pozbycie się przez kognitywistów reprezentacji jako zbędnego bagażu teoretycznego. Z drugiej jednak – także wśród tego rodzaju nowszych koncepcji znajdują się stanowiska bardziej „koncyliacyjne", zmierzające do takiego sformułowania reprezentacjonizmu, by był on zgodny z ideą systemu poznawczego jako ucieleśnionego i działającego w środowisku (por.: Grush 1997, 2004; Bickhard 2004a, 2004b, 2009; Anderson, Rosenberg 2008).

Dlaczego powracam tu do kwestii relacji między reprezentacjonizmem a nowymi trendami teoretycznymi w naukach kognitywnych? Otóż wydaje się, że zachodzi pewne powinowactwo między mechanistycznym spojrzeniem na rolę eksplanacyjną reprezentacji a teoriami zmierzającymi do pogodzenia reprezentacjonizmu z nowymi sposobami rozumienia natury systemu poznawczego. Związek ten widać choćby w poniższym cytacie:

> […] chcemy zasugerować, że naturaliście jest dostępna inna perspektywa patrzenia na te kwestie [dotyczące natury reprezentacji – P. G.], taka, która jest zgodna [...] z ekologicznym spojrzeniem na naturę organizmu. Nasze stanowisko głosi, że reprezentacje są tym, co reprezentacje robią (Anderson, Rosenberg 2008: 56).

W cytowanym artykule Michael Anderson i Gregg Rosenberg krytycznie odnoszą się do faktu, że teorie tworzone przez filozofów i kognitywistów są na ogół skoncentrowane na rozjaśnieniu relacji mię-

dzy nośnikiem reprezentacji a tym, co reprezentowane. Mówiąc słowami tych autorów, teorie te są na ogół „skupione na wejściu" (*input focused*), pozostawiając otwartą lub niedookreśloną kwestię tego, w jaki sposób organizm czy system poznawczy korzysta z reprezentacji. Diagnoza ta okazuje się bardzo zbliżona do tezy Marka Bickharda (2004a; 2004b; 2009), zgodnie z którą zdecydowana większość formułowanych na pograniczu filozofii i kognitywistyki koncepcji opiera się na myśleniu o reprezentacjach w kategoriach „kodowania" (*encoding*). Cel teorii reprezentacji rozwijanych w takim duchu stanowi objaśnienie tego, czym jest, czy też na czym się opiera bardzo ogólnie rozumiana „korespondencja" między nośnikiem a tym, co reprezentowane[10]. Podejściu skoncentrowanemu na kodowaniu towarzyszy wedle Bickharda idea, że reprezentacje są funkcjonalnie odrębne i niezależne w stosunku do mechanizmów odpowiadających za procesy motywacyjne oraz kontrolę działania. Jednakże ani Bickhard (2004a, 2004b, 2009), ani Anderson z Rosenbergiem (2008) nie postulują wyrugowania pojęcia reprezentacji z kognitywistyki. Pokazują oni raczej, że to nie sam reprezentacjonizm, ale właśnie tego rodzaju specyficzne, skupione na kodowaniu podejście jest niekompatybilne z postrzeganiem systemu poznawczego jako aktywnego w środowisku, cielesnego bytu. Zamiast odrzucać reprezentacjonizm, musimy raczej zmienić sposób myślenia o reprezentacjach. Powinniśmy mianowicie poszukać koncepcji bardziej skupiającej się na „wyjściu", a więc na tym, jaką funkcję reprezentacje spełniają *dla* systemu poznawczego. Powinniśmy potraktować je jako „narzędzia", dzięki którym systemy poznawcze osiągają określone cele. Reprezentacje są przede wszystkim tym, co „robią" w systemie. Teoria reprezentacji skrojonych na „potrzeby" realnych, działających systemów poznawczych musi rozpoznać i docenić ten fakt. Idea, że reprezentacje mają do spełnienia istotną eksplanacyjną rolę, nie jest z konieczności całkiem obca duchowi nowych podejść w ko-

[10] Można powiedzieć, że zgodnie z odróżnieniem wprowadzonym w poprzednim podrozdziale teorie reprezentacji jako kodowania koncentrują się wyłącznie na problemie treści. Teorie czyniące zadość postulatom Bickharda oraz Andersona z Rosenbergiem będą przede wszystkim teoriami funkcji reprezentacji.

gnitywistyce. Potrzeba jedynie pewnej reorientacji w myśleniu o naturze samych reprezentacji.

Opisane wyżej postulaty teoretyczne wiążą się z projektem realizowanym tu przeze mnie na dwa sposoby. Po pierwsze, zachodzi oczywiste podobieństwo między mechanistycznym spojrzeniem na rolę eksplanacyjną reprezentacji a ideą, że teoria reprezentacji powinna koncentrować się na funkcjonalnej roli odgrywanej przez reprezentacje w ramach systemu poznawczego. Można uznać, że mowa tu o dwóch ścieżkach – jedna wiedzie przez mechanicyzm, druga przez ucieleśnione i „działaniowe" postrzeganie natury systemu poznawczego – prowadzących do sformułowania tego samego problemu[11]. Poszukiwanie funkcjonalnej koncepcji reprezentacji

[11] Temu twierdzeniu można postanowić następujący zarzut. Bickhard (2004a, 2004b) oraz Anderson z Rosenbergiem (2008) mówią co prawda, że reprezentacje powinny w jakiś sposób „służyć" organizmowi czy systemowi poznawczemu. Nie formułują oni jednak swoich twierdzeń w kontekście jakiejś konkretnej teorii wyjaśniania w kognitywistyce. Ich stanowiska jako takie nie mówią nic o mechanizmach, ich komponentach i operacjach wykonywanych przez komponenty. Dlaczego mielibyśmy je zatem interpretować przez pryzmat mechanicyzmu? Otóż wydaje się, że postulaty tych badaczy można całkiem naturalnie odczytać za pomocą mechanistycznego modelu wyjaśniania. Kiedy autorzy ci mówią o „służeniu jako reprezentacja" albo „byciu reprezentacją dla organizmu", nie mają rzecz jasna na myśli sytuacji, w której reprezentacja „służy" komuś, tak jak służyć mu może pewna reprezentacja zewnętrzna – znak drogowy, wskazówka kompasu albo napis w języku naturalnym. Taka interpretacja generowałaby błąd homunkularny: wymagałaby postulowania istnienia *wewnątrz* systemu poznawczego „istot" obdarzonych zdolnościami interpretacyjnymi charakterystycznymi dla ludzi (por. Ramsey 2007: 26). W koncepcjach tego rodzaju nie chodzi o to, że system poznawczy jako taki interpretuje (wykorzystuje) wewnętrzną reprezentację, tak jak może on wykorzystywać reprezentację zewnętrzną. Bickhardowi oraz Andersonowi i Rosenbergowi chodzi raczej o sytuacje, w których pewien *podsystem organizmu* wykorzystuje jakiś inny podsystem w takiej roli. Takie rozumienie „służenia jako reprezentacja" jest zaś w sposób oczywisty bliskie mechanicyzmowi. Zamiast mówić o dwóch podsystemach, mechanicyzm będzie mówił o dwóch komponentach mechanizmu, które wchodzą w pewną interakcję (konkretnie interakcję polegającą na tym, że jeden komponent wykorzystuje drugi jako reprezentację). Rzecz jasna pojawia się tu problem, jak należy takie mechaniczne „korzystanie z reprezentacji" rozumieć. Pomijając póki co tę kwestię, na tym etapie wystarczy tylko podkreślić, iż Bickhard oraz Anderson z Rosenbergiem są zainteresowani *subosobowym*

nie jest więc motywowane jedynie mechanistycznym modelem wyjaśniania. Intelektualna potrzeba posiadania koncepcji funkcji reprezentacji naturalnie wynika też z innych, niezależnych trendów teoretycznych we współczesnej kognitywistyce. Pozwala to zarazem sądzić, że broniona przeze mnie dalej (w rozdziale 4) teoria mechanizmów reprezentacyjnych będzie nie tylko wartościowa jako teoria wyjaśniania reprezentacyjnego, ale potencjalnie może ona odegrać także pozytywną rolę w kontekście prób uzgodnienia reprezentacjonizmu z nurtami podkreślającymi znaczenie środowiska, działania i ucieleśnienia dla kognitywistyki.

Po drugie, należy postawić pytanie: skoro Bickhard i Anderson z Rosenbergiem poszukują funkcjonalnej koncepcji reprezentacji, to czy właśnie w ich teoriach nie powinniśmy znaleźć potencjalnego rozwiązania problemu mechanizmów reprezentacyjnych? Skoro autorzy ci stawiają podobny problem, to czy proponowane przez nich rozwiązania nie powinny okazać się użyteczne dla projektu realizowanego w tej pracy? W rzeczy samej niektóre elementy interakcyjnej teorii reprezentacji Bickharda (2004, 2009) zostaną przeze mnie wykorzystane w prezentowanej dalej koncepcji mechanizmów reprezentacyjnych. Nie wszystkie jednak nowsze teorie reprezentacji są równie przekonujące. Przyjrzyjmy się bliżej koncepcji reprezentacji jako przewodników działań (*action guidance theory of representation*) autorstwa Andersona i Rosenberga (2008). Wydaje się ona bowiem przykładem „pouczającego błędu": zawodzi jako koncepcja reprezentacji, jednak zawodzi w sposób, który dużo mówi o metodologicznych wyzwaniach stojących przed każdym teoretykiem zmierzającym do stworzenia funkcjonalnej koncepcji reprezentacji (por. także Gładziejewski 2015).

Według Andersona i Rosenberga (2008) rola reprezentacji mentalnych polega na kontrolowaniu działań. Być reprezentacją to grać rolę „przewodnika" działań organizmu (systemu poznawczego). Bardziej technicznie: egzemplarz (pewnego typu) *E* reprezentuje obiekt

rozumieniem reprezentacji mentalnych, czyli rozumieniem ich jako komponentów wewnętrznych mechanizmów poznawczych. (W rozdziale 5 postaram się pokazać, jak opozycję „osobowe–subosobowe" należy rozumieć w kontekście mechanistycznej koncepcji wyjaśniania).

O dla podmiotu *P* wtedy i tylko wtedy, gdy egzemplarze *E* są standardowo używane przez *P* po to, aby kierować działaniami *P* względem *O* (za: Anderson, Rosenberg 2008: 56–57)[12]. Należy wyraźnie zaznaczyć, że teoria Andersona i Rosenberga jest złożona i dość wyrafinowana pojęciowo. Autorzy ci dokładają wielu starań, by każdy termin, którym się posługują, został zdefiniowany precyzyjnie oraz w sposób unikający błędnego koła polegającego na analizowaniu pojęcia reprezentacji za pomocą kategorii presuponujących istnienie stanów reprezentacyjnych. Pełne przedstawienie koncepcji reprezentacji jako przewodników nie jest tu jednak konieczne. Jej ogólna idea pozostaje dość jasna i można ją sformułować w kategoriach zaczerpniętych z mechanicyzmu.

Wyobraźmy sobie, że opisujemy wewnętrzny mechanizm, który kontroluje czy steruje działaniami pewnego systemu. Przyjmijmy też, że mówimy o bardzo prostym systemie, którego cała aktywność polega na zbliżaniu się do pewnych obiektów (na przykład pożywienia czy partnerów reprodukcyjnych) oraz oddalaniu od innych (na przykład drapieżników). Na mechanizm sterujący naszym systemem składają się dwa komponenty. Przyjmijmy, że jeden z tych komponentów – nazwijmy go „M" – jest funkcjonalnie zaangażowany w kontrolowanie efektorów motorycznych naszego systemu. To, jakie sygnały M wysyła do efektorów, jest każdorazowo zależne od tego, w jakim stanie znajduje się inny komponent mechanizmu – nazwijmy ten komponent „R". Inaczej mówiąc, stan, w jakim znajduje się R, decyduje każdorazowo o tym, jakie rodzaje komend motorycznych zostaną w danych okolicznościach wygenerowane przez M. Załóżmy, iż w danej sytuacji R wpływa na M w taki sposób, że M generuje sekwencję ruchów polegających na zbliżeniu się systemu do pewnego obiektu O. Przyjmijmy ponadto, że nasz system jest zaprojektowany tak, iż przy zajściu podobnych okoliczności R standardowo sprawia, że M wysyła do efektorów komendy motoryczne służące zbliżaniu się do obiektów typu O. Tym samym R „przewodzi"

[12] Warto mieć na uwadze, że koncepcja Andersona i Rosenberga (2008) nie ogranicza klasy „działań", którymi przewodzą reprezentacje, do samych ruchów ciała. Procesy poznawcze czy inferencyjne także stanowią „działania" w przyjmowanym przez nich sensie tego terminu.

działaniami naszego systemu względem O. Robi to jednak, nie tyle bezpośrednio wpływając na sam system jako całość, lecz wchodząc w odpowiednie interakcje z M, czyli komponentem wewnętrznego mechanizmu tego systemu. Można powiedzieć, że zgodnie z (mechanistycznie zinterpretowaną) teorią reprezentacji jako przewodników działań w naszym przykładzie R *reprezentuje* O *dla* M. W ujęciu Anderson i Rosenberga mamy tu zatem tu do czynienia z mechanizmem reprezentacyjnym.

Na czym polega problem z teorią Andersona i Rosenberga? Otóż wydaje się, że można łatwo wskazać przykłady, w których to komponent mechanizmu przewodzi działaniami pewnego szerszego systemu (w sensie zakładanym przez teorię reprezentacji jako przewodników działań), jednak nie ma zarazem dobrych podstaw, by można było powiedzieć, iż pełni on funkcję reprezentacji. Jak należałoby chociażby podejść do przypadków, w których pewien wewnętrzny „przewodnik" działania pełni swoją funkcję na zasadzie odruchu? Weźmy pod uwagę stan neurofizjologiczny, w jakim znajdujemy się, kiedy nieopatrznie zbliżymy dłoń do płomienia, a który to (standardowo) sprawia, że dłoń tę natychmiast od płomienia odsuwamy. Stan ten przewodzi naszym działaniem względem płomienia, nie wiadomo jednak, na jakiej podstawie można mówić, że *reprezentuje* on płomień (lub kategorię: „zagrożenie"). Adaptywne dyspozycje behawioralne organizmów względem różnego rodzaju obiektów czy stanów rzeczy będą zawsze systematycznie zależne od stanów, w jakich znajdują się wewnętrzne struktury tych organizmów (stany te będą standardowo generować określone reakcje zachowaniowe na określone typy okoliczności). Nie ma takiego zachowania względem zewnętrznych okoliczności środowiskowych, które nie byłoby jakoś wewnętrznie zapośredniczone. Co sprawia, że pewne wewnętrzne struktury pośredniczące przyczynowo między okolicznościami środowiskowymi a działaniami są reprezentacjami, a inne nie? Omawiana teoria nie daje odpowiedzi na to pytanie. A może powinniśmy uznać, że wszelkie wewnętrzne struktury zaangażowane w kontrolę działania są w istocie reprezentacjami? Jeśli jednak Anderson i Rosenberg odpowiedzą twierdząco na to ostatnie pytanie, to ich koncepcja wydaje się trywializować pojęcie reprezenta-

cji, ponieważ każe nam interpretować jako reprezentacje struktury (komponenty mechanizmów), które *prima facie* takiej interpretacji się nie poddają. Co więcej, z koncepcji takiej wynikałoby, że antyreprezentacjonistyczne mechanistyczne wyjaśnianie działań systemów poznawczych jest w zasadzie niemożliwe: sam fakt, że coś (jakaś wewnętrzna struktura czy stan) systematycznie kontroluje działania systemu względem okoliczności środowiskowych, gwarantowałby temu czemuś status funkcjonalny reprezentacji. Nie sposób przyjąć takiego twierdzenia: antyreprezentacjonizm *stanowi* realną alternatywę dla reprezentacjonizmu.

Wydaje się zatem, że koncepcja reprezentacji jako przewodników działań jest po prostu zbyt gruboziarnista. Niewykluczone, że „przewodzenie działaniem" w sensie przyjmowanym przez Andersona i Rosenberga w istocie stanowi *genus* dla pojęcia reprezentacji w eksplanacyjnie wartościowym dla kognitywistyki sensie. Tym, czego brakuje teorii tych autorów, jest wskazanie *differentia specifica*, dzięki której można byłoby odróżnić mechanizmy, których komponenty przewodzą działaniem, *pełniąc funkcję reprezentacji*, od takich mechanizmów, których komponenty przewodzą co prawda działaniem, ale nie jako reprezentacje (na przykład na czysto asocjacyjnej lub odruchowej zasadzie).

Biorąc powyższe ustalenia pod uwagę, odróżnijmy zatem dwa pytania:

1. Na czym polega pełnienie roli reprezentacji (służenie jako reprezentacja) przez komponent mechanizmu?
2. Na jakiej podstawie możemy orzekać, że posiadanie określonych własności funkcjonalnych przez komponent mechanizmu gwarantuje mu status reprezentacji (nośnika reprezentacji)?

W omawianym artykule Anderson i Rosenberg proponują określoną odpowiedź na pytanie 1. Ich propozycja wydaje się jednak nietrafiona. Ogólna idea reprezentacji jako czegoś, co kieruje działaniami, jest wstępnie zrozumiała i wiarygodna, ale przy bliższym spojrzeniu okazuje się podatna na kontrprzykłady. Nie wiadomo, dlaczego właśnie ten konkretny, proponowany przez autorów profil funkcjo-

nalny należałoby uznać za charakteryzujący reprezentacje. Można postawić hipotezę, że przyczyna takiego stanu rzeczy leży w fakcie, iż Anderson i Rosenberg nie postawili sobie otwarcie pytania 2. Nie podjęli się oni próby pokazania, na jakich dokładnie podstawach wskazane przez nich własności funkcjonalne reprezentacji zasługują na miano własności cechujących wewnętrzne *reprezentacje*. Pewne opisy funkcjonalne komponentów mechanizmu będą odnosiły się do ról rzeczywiście reprezentacyjnych (komponenty takie będą pełniły funkcję reprezentacji), a inne – do ról, które takimi nie są (komponenty takie nie będą pełniły funkcji reprezentacji). Tym, czego potrzebujemy, poszukując teorii mechanizmów reprezentacyjnych, jest wyrażona *explicite* metoda, za pomocą której moglibyśmy – w nieарbitralny, dobrze uzasadniony sposób – odróżnić jedne przypadki od drugich. Systematyczne podjęcie tej metodologicznej kwestii pozwoli nam odpowiednio ukierunkować poszukiwanie koncepcji mechanizmów reprezentacyjnych i uniknąć mielizn takich jak ta, na którą natrafia teoria Andersona i Rosenberga[13].

3.2.2. Czym jest funkcja reprezentowania? Wymóg opisu zadań i metoda Ramseya

Jak się okazuje, bardzo użyteczna propozycja dotycząca metodologicznego remedium na zaznaczony powyżej problem została sformułowana przez Williama Ramseya w jego pracy *Representation Reconsidered* (2007: 1–37). Remedium to będę tu skrótowo nazywać „metodą Ramseya". Zanim przejdę do jej opisu, warto naszkicować szerszy projekt realizowany przez tego autora we wspomnianej pracy.

Ramsey wychodzi od podejrzenia, że status reprezentacji mentalnych jako narzędzia eksplanacyjnego kognitywistyki przypomina obecnie status pojęcia sfer niebieskich po rewolucji kopernikańskiej w astronomii. W systemie ptolemejskim sfery niebieskie odgrywały istotną rolę eksplanacyjną: to na nich były podwieszone ciała niebie-

[13] Jak zobaczymy dalej, problemami tego rodzaju jest obarczona nie tylko ta konkretna teoria, ale także wiele innych kognitywistycznych koncepcji reprezentacji.

skie obracające się wokół nieruchomej Ziemi. Powstanie nowej, heliocentrycznej kosmologii sprawiło, że sfery niebieskie stały się eksplanacyjnie zbędne. Nieruchome gwiazdy nie „potrzebowały" już sfer, po których miałyby się poruszać. Jak zwraca uwagę Ramsey, mimo takiego stanu rzeczy Mikołaj Kopernik, Tycho Brahe oraz Johannes Kepler nie wyeliminowali całkowicie pojęcia sfer niebieskich z astronomii (doprowadziło do tego dopiero powstanie kartezjańskiej koncepcji przestrzeni). Utrzymywanie pojęcia sfer niebieskich w astronomii bezpośrednio po rewolucji Kopernika nie było już jednak motywowane rzeczywistymi, merytorycznymi potrzebami eksplanacyjnymi. Jak zatem wytłumaczyć ten swoisty teoretyczny konserwatyzm? Wydaje się wiarygodne, że sfery niebieskie pozostały obecne w nowej, heliocentrycznej teorii w wyniku oddziaływania pozamerytorycznych czynników o charakterze historycznym, społeczno-kulturowym czy też na podstawie intuicji wywodzących się z „potocznego" rozumienia zjawisk fizycznych.

Wedle Ramseya (2007: 1–37) sytuacja reprezentacji mentalnych na obecnym etapie historycznego rozwoju kognitywistyki przypomina do pewnego stopnia sytuację sfer niebieskich po rewolucji kopernikańskiej. Reprezentacje do pewnego momentu były co prawda eksplanacyjnie wartościowe dla kognitywistyki. W pierwszej dekadzie XXI wieku tak już jednak nie jest. Ramsey twierdzi, że pojęcie reprezentacji nie pełni obecnie żadnej rzeczywistej, wartościowej roli eksplanacyjnej w naukach kognitywnych. Kognitywiści (oraz wykorzystujący kognitywistykę filozofowie) nadal posługują się terminem „reprezentacja", jednak kiedy przyjrzymy się desygnatom tego terminu, okaże się, że nie można ich zasadnie uznać za reprezentacje. Termin „reprezentacja" widnieje w nowych wyjaśnieniach działania systemu poznawczego zapewne z powodów pozamerytorycznych: w wyniku obawy przed powrotem behawioryzmu, na podstawie założeń o naturze umysłu wywodzących się z psychologii potocznej, czy w wyniku jeszcze innych czynników. Niezależnie jednak od tego, co jest przyczyną tego stanu rzeczy – zgodnie z centralną tezą Ramseya – podobnie jak sfery niebieskie po rewolucji Kopernika, tak i reprezentacje mentalne utrzymują swoje miejsce w nauce przez swoisty pojęciowy „zastój", nie spełniając żadnych rzeczywi-

stych ról eksplanacyjnych. Wedle tego autora chociaż nie zdajemy sobie (jeszcze) z tego sprawy, to żyjemy w erze postreprezentacjonistycznej kognitywistyki już od czasu powstania koneksjonizmu.

Na tym wreszcie etapie rozważania Ramseya zazębiają się z tymi prowadzonymi w tej książce. Na jakiej podstawie bowiem autor ten broni swojego mocnego, rewizjonistycznego stanowiska? Otóż odwołuje się on do ogólnego założenia, że status eksplanacyjny reprezentacji wiąże się ściśle z funkcjami pełnionymi przez reprezentacje w ramach systemu poznawczego (Ramsey 2007: 28)[14]. Pełnoprawnie reprezentacyjne wyjaśnienia zjawisk poznawczych odwołują się do wewnętrznych struktur, których funkcją jest reprezentowanie czegoś. Centralny punkt argumentacji Ramseya stanowi tak zwany wymóg opisu zadań (*job description challenge*). Aby sprostać temu wymogowi, należy pokazać, w jakim sensie – czy też w jaki sposób – pewien wewnętrzny stan lub struktura (postulowana jako reprezentacja) pełni funkcję reprezentacji w ramach systemu poznawczego[15].

---

[14] Ramsey przyjmuje to założenie bez odwoływania się *explicite* do określonej koncepcji wyjaśniania naukowego. Teoria mechanistyczna pokazuje jednak, że (jak również – dlaczego) założenie to jest trafne.

[15] Ramsey (2007: 11–12) wprowadza bardzo użyteczne rozróżnienie na sytuacje, w których coś odgrywa rolę eksplanacyjną *simpliciter*, oraz takie, w których coś odgrywa rolę eksplanacyjną *jako reprezentacja*. Dystynkcję tę ilustruje podany przez autora (Ramsey 2007: 11–12) przykład hipotetycznej sytuacji, w której pewien badacz proponuje reprezentacyjną teorię powstawania chorób. Przy bliższym spojrzeniu okazuje się jednak, że teoretyk ten przypisuje postulowanym przez siebie (rzekomym) „reprezentacjom" role pełnione przez „zwykłe" czynniki biologiczne (bakterie i wirusy) przyczyniające się do powstawania chorób. Załóżmy, że proponowana przez tego badacza koncepcja powstawania chorób jest prawdziwa. Odwołuje się ona do czynników rzeczywiście chorobogennych. Czynniki nazywane przez nią „reprezentacjami" rzeczywiście wyjaśniają genezę chorób. Nie oznacza to jednak, że wyjaśniają powstawanie chorób jako reprezentacje. Omawiana teoria jest poprawna, jednak nie wiadomo, w jakim sensie miałaby ona być *reprezentacyjna*. Przecież role funkcjonalne, które przypisuje ona czynnikom chorobogennym, nie polegają na reprezentowaniu czegoś. Należy więc przyjąć, że w teorii tej pojęcie reprezentacji nie spełnia żadnej roli, że teoria ta, choć poprawna, jest jedynie pozornie reprezentacyjna. Analogicznie, Ramseyowi nie chodzi o to, że byty postulowane jako reprezentacje przez współczesne teorie czy modele systemu poznawczego nie spełniają istotnych ról eksplanacyjnych. Wymóg opisu zadań ma jedynie pomóc zdiagnozować, czy

Należy pokazać, w jakim sensie „zadania" pełnione przez pewien byt postulowany przez teorię kognitywistyczną są „reprezentacyjne". Jeśli pewien teoretyk posługuje się terminem „reprezentacja", a jednak przy bliższym spojrzeniu okazuje się, że desygnaty tego pojęcia nie odgrywają w żadnym sensie roli reprezentacji, to omawiana teoria jest reprezentacyjna jedynie pozornie. Zasadnicze znaczenie dla argumentacji Ramseya ma konstatacja, że kiedy przyjrzymy się najlepszym rzeczywistym, dostępnym koncepcjom działania systemu poznawczego, to okaże się, że nawet jeśli te koncepcje powołują się na reprezentacje jako ważne narzędzie eksplanacyjne, to robią to w sposób, który nie czyni zadość wymogowi opisu zadań. Koncepcje te mogłyby równie dobrze obyć się bez reprezentacyjnej terminologii, nie tracąc swojej wartości eksplanacyjnej. Inaczej mówiąc, są one jedynie pozornie reprezentacyjne.

Zwróćmy uwagę, że istnieje związek między wymogiem opisu zadań a postawionym wcześniej pytaniem 2: „Na jakiej podstawie możemy orzekać, że posiadanie określonych własności funkcjonalnych przez komponent mechanizmu gwarantuje mu status reprezentacji (nośnika reprezentacji)?". Ramsey potrzebuje bowiem jakiegoś metodologicznego narzędzia, które pozwoliłoby mu odróżnić uprawnione odwołania do reprezentacji mentalnych od tych nieuprawnionych. Potrzebuje on jakiegoś kryterium pozwalającego ocenić, czy dany rodzaj struktury postulowanej jako reprezentacja rzeczywiście czyni zadość wymogowi opisu zadań. Niezbędna okazuje się tu więc metoda pozwalająca na odróżnienie przypadków, w których jakiś funkcjonalnie scharakteryzowany komponent systemu poznawczego jest zasadnie uznawany za reprezentację, od takich, w których taka jego kategoryzacja jest bezpodstawna. Zaproponowanie takiej procedury równa się udzieleniu odpowiedzi na pytanie 2. Metoda Ramseya ma służyć takiemu właśnie celowi. Na czym ona zatem polega? Choć sam Ramsey (2007) nie prezentuje jej *explicite* w taki sposób, to można tę metodę scharakteryzować ogólnie jako procedurę składającą się z trzech kroków:

---

rzeczywiście spełniają one te role jako reprezentacje (to znaczy czy byty te zajmują się rzeczywiście reprezentowaniem).

a) Rekonstrukcji przednaukowego pojęcia (pojęć) reprezentacji w kategoriach funkcjonalnych.
b) Rekonstrukcji pojęć reprezentacji obecnych w teoriach, modelach i wyjaśnieniach kognitywistycznych.
c) Określenia, czy „reprezentacje" desygnowane przez pojęcia wymienione w (b) funkcjonują w sposób na tyle zbliżony do sposobu funkcjonowania reprezentacji w sensie przednaukowym, aby można im było przypisać rolę reprezentowania.

Przyjrzymy się teraz bliżej kolejnym z tych kroków.

Ad a) Rekonstrukcja przednaukowego pojęcia (pojęć) reprezentacji w kategoriach funkcjonalnych

Pierwszy krok metody Ramseya (2007: 8–24) wiąże się ściśle z założeniem, że rodowód kognitywistycznego pojęcia reprezentacji wykracza poza naukę. Pojęcie reprezentacji mentalnych stosowane w kognitywistyce różni się pod tym względem chociażby od biologicznego pojęcia genu. To ostatnie jest czysto technicznym narzędziem pojęciowym, które zostało sformułowane w celu spełnienia potrzeb eksplanacyjnych powstałych w ramach biologii. W przeciwieństwie do pojęcia genu – pojęcie reprezentacji ma źródła w *przednaukowych* sposobach konceptualizowania świata przez ludzi. W omawianej koncepcji nie jest to jednak tylko czysto opisowa teza dotycząca źródeł pojęcia reprezentacji w kognitywistyce. Ramsey uznaje dodatkowo, że przednaukowe sposoby posługiwania się tym pojęciem powinny w istotny sposób *ukierunkowywać* i *ograniczać* to, jak posługują się nim naukowcy. Reprezentacje są przydatne w kognitywistyce o tyle, o ile system poznawczy działa na podstawie stanów czy struktur w jakimś nietrywialnym stopniu przypominających reprezentacje w przednaukowym sensie. Czysto techniczne definicje funkcjonalne reprezentacji formułowane przez kognitywistów – w rodzaju tej proponowanej przez Andersona i Rosenberga (2008) – uderzają nas jako zasadniczo niedoskonałe i podatne na kontrprzykłady właśnie dlatego, że kłócą się one z przednaukowym rozumieniem reprezentacji. Kontrprzykłady dla tych definicji to sytuacje, w których pewna wewnętrzna struktura spełnia podaną

przez autorów definicję, ale jest kompletnie nieoczywiste, dlaczego mielibyśmy uznać pełnioną przez nią w systemie czy mechanizmie rolę za rolę polegającą na reprezentowaniu czegoś. Co więcej, wydaje się jasne, że decyzja o zastosowaniu tu terminologii reprezentacyjnej jest w danym przypadku bezzasadna[16]. Zgodnie z propozycją Ramseya (2007: 8–11) za taki stan rzeczy każdorazowo odpowiada fakt, że rola pełniona przez daną strukturę oddala się od przednaukowego sposobu rozumienia reprezentacji na tyle, że nie ma w niej już nic rozpoznawalnie „reprezentacyjnego". Jeśli uznanie systemu poznawczego za (w jakimś zakresie) reprezentacyjny ma stanowić ważną i bogatą w konsekwencje tezę, to sposób posługiwania się terminem „reprezentacja" musi być czymś ograniczony. Biorąc pod uwagę przednaukowe źródła tego pojęcia, owo ograniczenie powinno pochodzić spoza samej kognitywistyki.

Zauważmy, iż kwestia, o której tu mowa, nie jest czysto terminologiczna. Ktoś mógłby bowiem powiedzieć, że termin „reprezentacja" w jego naukowym zastosowaniu ma charakter techniczny i to sami kognitywiści – w zależności od swoich potrzeb teoretycznych i eksplanacyjnych – ustalają sposób, w jaki będą go używać. Dlaczego mieliby się oni przejmować potocznymi pojęciami reprezentacji oraz ich filozoficznymi analizami czy rekonstrukcjami? Ramsey (2007: 11–14) wymienia kilka takich powodów. Po pierwsze, całko-

[16] Oprócz wspomnianych już kontrprzykładów dla teorii Andersona i Rosenberga warto przytoczyć tu przypadek wskazany przez samego Ramseya (2007: 9). Przywołuje on następującą definicję reprezentowania („desygnowania") zaczerpniętą z pracy Allena Newella (1980: 156): „Jednostka *X* desygnuje jednostkę *Y* relatywnie do procesu *P*, jeśli jest tak, że kiedy *X* stanowi wejście [*input*] dla *P*, to przebieg tego procesu jest zależny od *Y*". Raz jeszcze okazuje się, że nietrudno wskazać przypadki, w których definicja ta jest spełniona, ale których nie sposób uznać za przypadki reprezentowania. Kontrprzykład podany przez samego Ramseya (2007) to sytuacja, w której czyjeś procesy trawienne przyjmują zimny napój na „wejściu", a sposób ich przebiegu zależy między innymi od tego, czy dana osoba coś wcześniej zjadła. Uznanie, że rola wypitego napoju polega na reprezentowaniu czegoś, jest nienaturalne i zbędne eksplanacyjnie. Podkreślmy, że nie chodzi o to, iż wypicie napoju nie będzie tu pełniło żadnej roli eksplanacyjnej. Problem polega na tym, że nie istnieją żadne dobre podstawy do tego, by uznać, że rola ta będzie polegać na reprezentowaniu faktu, iż ktoś wcześniej zjadł (lub nie) posiłek.

wite zignorowanie przednaukowej praktyki pojęciowej zagrażałoby sytuacją, w której termin reprezentacja może być stosowany przez kognitywistów w sposób w zasadzie dowolny. Stosowanie terminologii reprezentacyjnej w taki „techniczny" sposób łatwo doprowadza do sytuacji, w której nie da się określić, w jakim sensie czy też na jakiej podstawie dany byt postulowany (i określany jako „reprezentacja") przez teorię pełni swoją rolę eksplanacyjną jako reprezentacja (por. przypis 15). Potrzebujemy przednaukowego pojęcia reprezentacji jako zewnętrznej instancji czy pojęciowej „kotwicy" pozwalającej nam uniknąć takiej dowolności w stosowaniu terminologii reprezentacyjnej w kognitywistyce. Po drugie, określenie czegoś „reprezentacją" wiąże się z przyjęciem szeregu założeń i oczekiwań dotyczących jego natury, które to założenia i oczekiwania mają swoje źródło właśnie w przednaukowym sposobie rozumienia reprezentacji. Na przykład przyjmujemy naturalnie, że jedną z cech reprezentacji jest możliwość reprezentowania czegoś błędnie. Idea ta ma swoje źródła w „potocznym" rozumieniu reprezentacji. Właśnie dlatego uważamy, że dobra naukowa teoria odwołująca się do pojęcia reprezentacji powinna umożliwiać pokazanie, że oraz jak możliwe są reprezentacje błędne. „Technicznego" pojęcia reprezentacji, zgodnie z którym błędy są niemożliwe, w istocie nie powinniśmy uznawać w ogóle za pojęcie reprezentacji. Po trzecie, Ramsey zauważa, że krytykowane przez niego, nieuprawnione używanie terminu „reprezentacja" to nie czysto terminologiczna pomyłka, polegająca na tym, że pewną rzecz nazywa się za pomocą nieodpowiedniego terminu, jak wtedy, gdy mylnie nazwiemy psa „kotem". Chodzi mu raczej o sytuacje, w których tak nieodpowiednio nazwanej rzeczy są dodatkowo przypisywane własności posiadane przez rzeczywiste desygnaty danego terminu – jak w sytuacji, gdy mylnie uznamy, że psy są rodzajem kotów i mają „kocie" własności. Inaczej mówiąc, nie chodzi tu o pomyłkę terminologiczną, ale o rzeczywiste zamieszanie pojęciowe, gdzie stany czy struktury niestanowiące reprezentacji – nieposiadające charakterystycznych dla reprezentacji własności – są za nie uznawane. Po czwarte wreszcie, jeśli nie uznamy, że uprawnione posługiwanie się pojęciem reprezentacji w kognitywistyce jest w jakiś sposób ograniczane jej potocznym rozumieniem, to stwarzamy sytu-

ację, w której istotne teoretyczne problemy kognitywistyki mogą być rozwiązywane przez zmianę konwencji językowych. Merytoryczne pytanie o to, czy (lub w jakim zakresie) system poznawczy jest reprezentacyjny, mogłoby zostać „rozwiązane" przez czysto konwencjonalną decyzję terminologiczną, odpowiednio liberalizującą sposób posługiwania się przez naukowców terminem „reprezentacja". (Jest to analogiczne do sytuacji, w której problem, czy inne naczelne niż ludzie są zdolne posługiwać się językiem, ktoś próbowałby „rozwiązać", definiując „język" jako dowolny system komunikacji). Uczynienie przednaukowego pojęcia reprezentacji „kotwicą" dla jego wariantu naukowego blokuje taką możliwość.

Należy w tym kontekście równie wyraźnie zaznaczyć, że na gruncie podejścia prezentowanego przez Ramseya nie twierdzi się, iż przednaukowe pojęcie stanowi normę czy wzorzec, wyznaczający jednoznacznie i całkowicie sposób posługiwania się pojęciem reprezentacji przez kognitywistów. Niewykluczone, że reprezentacje przez nich odkrywane będą posiadać własności, których nie mają reprezentacje w sensie przednaukowym, własności, których istnienia obecnie nawet nie podejrzewamy i których nie potrafimy skonceptualizować. Ramseyowi nie chodzi o to, by naukowcy biernie, bez modyfikacji przejęli potoczne pojęcie reprezentacji. Idea wyzwania z opisu obowiązków wspiera się jedynie na – jak się wydaje, dość niewinnym: tak filozoficznie, jak naukowo – założeniu, iż aby w ogóle racjonalnie zakwalifikować coś jako reprezentację, należy pokazać, że (oraz na jakiej zasadzie) struktura postulowana jako reprezentacja posiada treść (warunki poprawności), może być fałszywa, pełni funkcję poznawczego „zastępstwa" przedmiotu reprezentacji i tak dalej. Ramsey zauważa jedynie, że tego rodzaju podstawowe własności decydujące o samym byciu reprezentacją mają pozanaukowy rodowód.

Z wszystkich wymienionych względów pierwszy krok metody Ramseya polega na rekonstrukcji przednaukowego sposobu postrzegania reprezentacji (por. tabela 2). Ramsey (2007) w praktyce nie odwołuje się bezpośrednio do „potocznych" pojęć reprezentacji, ale przede wszystkim do ich dostępnych filozoficznych analiz czy eksplikacji. Wedle tego autora na podstawie dostępnej literatury mo-

Tabela 2. Williama Ramseya klasyfikacja przednaukowych pojęć reprezentacji

Rodzaj pojęcia reprezentacji	Przednaukowe pojęcie reprezentacji mentalnych	Przednaukowe pojęcie reprezentacji pozamentalnych
Czym są reprezentacje?	Postawy propozycjonalne, czyli stany intencjonalne, do których odwołują się wyjaśnienia i predykacje sformułowane w ramach psychologii potocznej (na przykład: przekonania, pragnienia, nadzieje, intencje, obawy). Ich podmiotami są istoty, których działania są przewidywane i wyjaśniane za pomocą psychologii potocznej.	Zewnętrzne struktury wykorzystywane w roli reprezentacji przez podmioty intencjonalne: Reprezentacje ikoniczne, oparte na podobieństwie zachodzącym między nośnikiem reprezentacji a tym, co reprezentowane (na przykład: mapy, obrazy, fotografie). Reprezentacje indeksowe, oparte na systematycznej współzmienności zachodzącej między nośnikiem reprezentacji a tym, co reprezentowane (na przykład: termometr, kompas, prędkościomierz). Reprezentacje konwencjonalne, oparte na konwencjonalnych regułach interpretacji (na przykład: zdania sformułowane w językach naturalnych, niektóre znaki drogowe).
Na czym polega rola funkcjonalna reprezentacji?	Reprezentacje te (1) mają niewywiedzioną treść intencjonalną oraz (2) treść ta determinuje odgrywane przez nie role przyczynowe/funkcjonalne (w stosunku do stanów percepcyjnych, innych postaw propozycjonalnych oraz działań).	Reprezentacje te są wykorzystywane przez podmiot w celu informowania go czy udostępniania mu wiedzy na temat tego, co reprezentowane (stanu rzeczy, obiektu czy zdarzenia). Ich rola polega na „zastępowaniu" (stand-in) dla podmiotu tego, co reprezentowane; umożliwiają one praktyczne i poznawcze orientowanie się względem tego, co reprezentowane.

Na podstawie: Ramsey 2007: 14–24.

żemy orzec, że istnieją *de facto* dwa przednaukowe pojęcia reprezentacji. Pierwsze z nich to pojęcie reprezentacji *mentalnych*, wywodzące się z psychologii potocznej. Jego rekonstrukcję opiera Ramsey na rozwijanych przez filozofów w ostatnich dziesięcioleciach rozważaniach nad naturą psychologii potocznej i wchodzących w jej skład predykatów mentalnych. Zgodnie z tym rozumieniem reprezentacje to postawy propozycjonalne. Mają one niewywiedzioną treść intencjonalną (jej posiadanie nie jest pochodne względem czy zależne od aktu interpretacji dokonywanego przez podmiot intencjonalny) oraz grają role przyczynowe, które są determinowane przez treść. Drugie przednaukowe pojęcie wyróżniane przez Ramseya to pojęcie reprezentacji *pozamentalnych*. Jego rekonstrukcję opiera Ramsey z kolei przede wszystkim na pewnych klasycznych założeniach semiotyki Charlesa Peirce'a (por.: Peirce 1997; Short 2007; Atkin 2010). Chodzi w tym przypadku o zewnętrzne, artefaktualne lub naturalne reprezentacje, które swój status reprezentacji zawdzięczają temu, że są wykorzystywane w roli reprezentacji przez podmioty intencjonalne. Podmioty te traktują czy interpretują je jako reprezentacje określonych stanów rzeczy, obiektów czy zdarzeń. Reprezentacje pozamentalne zapośredniczają dostęp poznawczy podmiotu do tego, co reprezentowane; inaczej mówiąc, zastępują (*stand-in*) one dla podmiotu to, co reprezentowane. Wyniki rekonstrukcji Ramseya są przedstawione w tabeli 2.

Zanim przejdziemy do omówienia drugiego kroku przedstawianej procedury, warto zapytać, co w koncepcji Ramseya robi rekonstrukcja potocznego pojęcia reprezentacji pozamentalnych (zewnętrznych), skoro przedmiotem jego zainteresowania pozostaje rola eksplanacyjna reprezentacji mentalnych (wewnętrznych)? Otóż sposób rozumienia reprezentacji mentalnych przez kognitywistów jest często wzorowany na potocznym czy przednaukowym pojęciu reprezentacji pozamentalnych. Jak zobaczymy w dalszej części tej książki, reprezentacje w sensie postulowanym przez teorie z zakresu kognitywistyki okazują się często „zinternalizowanymi", mechanicznymi wersjami reprezentacji zewnętrznych.

Ad b) Rekonstrukcja pojęć reprezentacji obecnych w teoriach, modelach i wyjaśnieniach kognitywistycznych

Także drugi krok metody Ramseya (2007) polega na rekonstrukcji pojęć. W tym przypadku przedmiotem analizy jest naukowa, kognitywistyczna praktyka pojęciowa. To właśnie reprezentacje w naukowym sensie mają być „kandydatami" do sprostania wymogowi opisu zadań. Jacy to jednak kandydaci? Aby odpowiedzieć na to pytanie, należy ustalić, w jaki sposób pojęciem reprezentacji posługują się sami kognitywiści, gdzie reprezentacje są rozumiane jako byty postulowane przez teorie i modele działania systemu poznawczego. Jest to rzecz jasna zupełnie nietrywialne wyzwanie. Przedstawiciele nauk kognitywnych używają terminu „reprezentacja" na wiele różnych sposobów i w wielu znaczeniach, które tylko czasem są wyrażane *explicite*. Aby ułożyć ten chaotyczny gąszcz w nadającą się do opracowania całość, Ramsey podejmuje się zadania polegającego na wyróżnieniu kilku nadrzędnych kategorii stanów czy struktur uznawanych za reprezentacje zgodnie z pojęciową praktyką kognitywistów. Co należy zaznaczyć, zgodnie z planem autora za kryterium wyróżniania tych kategorii reprezentacji mają służyć ich role funkcjonalne. Poszczególne byty postulowane jako reprezentacje w ramach teorii czy modeli kognitywistycznych są w takiej perspektywie przypisywane do określonych kategorii ze względu na funkcje pełnione przez nie w systemie poznawczym. Ramsey wyróżnia w ten sposób pięć pojęć reprezentacji obecnych w kognitywistyce:

> **Reprezentacje jako postawy propozycjonalne.** W tym przypadku reprezentacje w sensie naukowym mają być „znaturalizowaną" wersją postaw propozycjonalnych. Zgodnie z tym pojęciem reprezentacje w sensie kognitywistycznym są wewnętrznymi stanami posiadającymi niewywiedzioną treść intencjonalną[17] oraz pełniącymi na podstawie tej treści role funkcjonalne, które odpowiadają rolom pełnionym przez postawy propozycjonalne.

[17] Jednocześnie treści stanów tego rodzaju mają odpowiadać treściom przypisywanym zwykle postawom propozycjonalnym.

Najbardziej wpływową teorią wykorzystującą to pojęcie reprezentacji jest koncepcja Jerry'ego Fodora (1975; 1987; 2001).

IO-reprezentacje (*input-output representations*). Operowanie tym pojęciem reprezentacji polega na przypisywaniu treści symbolom biorącym udział w procesach obliczeniowych dokonujących się w systemie poznawczym. Dzięki takiej atrybucji treści możliwe staje się zrozumienie zarówno tego, (1) jaka operacja obliczeniowa jest wykonywana w systemie, jak i tego, (2) na czym polegają obliczeniowe kroki pośredniczące między symbolami znajdującymi się na wejściu oraz wyjściu danego procesu obliczeniowego[18]. Jest to pojęcie stosowane rutynowo w obliczeniowych modelach działania systemu poznawczego, w których obliczanie rozumie się „klasycznie", czyli jako obliczanie symboliczne (to znaczy jako przetwarzanie wewnętrznych symboli na podstawie ich formalnych czy syntaktycznych własności).

S-reprezentacje. Zgodnie z tym pojęciem reprezentacje to wewnętrzne symulacje lub modele reprezentowanych obiektów, stanów rzeczy czy procesów (reprezentacjami mogą być też nazywane elementy składowe takich modeli). Status modeli zawdzięczają one zachodzeniu strukturalnego podobieństwa między nośnikiem reprezentacji a tym, co reprezentowane. Z pojęcia S-reprezentacji korzystały na przykład niektóre klasyczne, symboliczne modele obliczeniowe czynności poznawczych (jak chociażby SOAR; por. Laird, Newell, Rosenbloom 1987), a także psychologiczna teoria modeli mentalnych Philipa Johnsona-Lairda (1983).

Reprezentacje receptorowe. Zgodnie z tym pojęciem reprezentacje pełnią w systemie poznawczym rolę receptorów czy

[18] Na przykład, możemy zidentyfikować operację wykonywaną przez system (powiedzmy, jako mnożenie „pięć razy pięć"), dzięki zinterpretowaniu symboli na wejściu i wyjściu jako reprezentujących liczby. Na podobnej zasadzie możemy zrozumieć składowe operacje obliczeniowe, które pośredniczą między symbolami znajdującymi się na wejściu oraz wyjściu badanego procesu (na przykład potrafimy stwierdzić, że mamy do czynienia z pięciokrotnym wykonaniem operacji polegającej na dodaniu do siebie liczby pięć). Bez przypisania treści symbolom – nie zrozumiemy, jaką operację wykonuje pewien system oraz jak on ją wykonuje.

detektorów. Pełnienie tej roli opiera się na występowaniu systematycznej współzmienności między stanami, w jakich znajduje się nośnik reprezentacji, a stanami tego, co podlega reprezentowaniu. Pojęcie to stosuje się w neuronauce poznawczej, chociażby w kontekście „receptorowych" własności neuronów znajdujących się w korze wzrokowej u różnych gatunków organizmów (por. np.: Lettvin, Maturana, McCulloch, Pitts 1959; Logothetis, Sheinberg 1996; Quiroga, Reddy et al. 2005).

Reprezentacje ukryte. Reprezentacje w tym rozumieniu nie są zlokalizowane w określonym komponencie systemu poznawczego (komponencie mechanizmu poznawczego). Co więcej, reprezentacje ukryte nie są w ogóle *explicite* kodowane w systemie. Zamiast tego – są one przypisywane systemowi na podstawie jego dyspozycji do podejmowania określonych działań przy zajściu określonych okoliczności. Pojęcie reprezentacji ukrytych bywa stosowane między innymi w kontekście modeli koneksjonistycznych funkcji poznawczych (por. np.: Churchland 1989; Clark 1993; Rogers, McClelland 2004). Operowanie tym pojęciem polega w tym kontekście na przypisywaniu sieci koneksyjnej ukrytych reprezentacji wtedy, gdy układ wag w pośredniczących warstwach neuronów sprawia, że sieć ma dyspozycje do określonych wzorców zachowań/reakcji względem różnych kategorii obiektów czy stanów rzeczy.

Powyższa lista ujmuje rzecz jasna jedynie zarys rekonstrukcji pojęciowej przeprwoadzonej przez Ramseya. Niektóre z wymienionych wyżej pojęć reprezentacji – receptorowe, ukryte oraz to odwołujące się do S-reprezentacji – zostaną szerzej omówione w dalszej części tej pracy. Póki co należy jednak zwrócić uwagę na jeden ważny aspekt Ramseya klasyfikacji kognitywistycznych pojęć reprezentacji. Wedle tego autora istnieje pewien związek między poszczególnymi naukowymi *pojęciami* reprezentacji a ogólnymi *podejściami* do modelowania i wyjaśniania działania systemu poznawczego (Ramsey 2007: 1–4, 203–222). Mówiąc dokładniej, kognitywistyczne pojęcia reprezentacji dzielą się wedle Ramseya na: (1) takie, które spełniały rolę eksplanacyjną w ramach klasycznego, obliczeniowo-

-symbolicznego podejścia, oraz (2) takie, które znalazły w kognitywistyce zastosowanie dopiero wraz z wyłonieniem się podejść nowszych, uznawanych za alternatywne wobec podejścia klasycznego. Według Ramseya pierwsze trzy wymienione wyżej pojęcia reprezentacji – czyli pojęcia reprezentacji jako postaw propozycjonalnych, IO-reprezentacji oraz S-reprezentacji – należą do (1). Powstały one i rozpowszechniły się w ramach klasycznego podejścia do rozumienia natury procesów poznawczych. Ponadto ich użyteczność zawężała się według Ramseya jedynie do teorii powstałych w ramach tego podejścia. Dwa ostatnie wymienione pojęcia reprezentacji – pojęcie reprezentacji receptorowych i ukrytych – należą według Ramseya do (2). Znalazły one użytek w ramach nowszych podejść w kognitywistyce, zwłaszcza w koneksjonizmie i neuronauce poznawczej.

Taki podział nie pełni w koncepcji Ramseya wyłącznie klasyfikacyjnej roli, lecz jest kluczowy dla realizacji nadrzędnego celu *Representation Reconsidered*, polegającego na dostarczeniu diagnozy o rzekomo niereprezentacyjnym (czy może „krypto-antyreprezentacjonistycznym") charakterze kognitywistyki na jej obecnym etapie rozwoju. Sądzę zarazem, że właśnie tu Ramsey popełnił w swojej rekonstrukcji zasadniczą i bogatą w konsekwencje pomyłkę. Chodzi o dokonane przez niego przypisanie pojęcia S-reprezentacji do (wyłącznie) podejścia klasycznego, czyli (1). Jak zobaczymy w rozdziale 4 (podrozdział 4.3), rozpoznanie tej pomyłki stanowi klucz do odrzucenia Ramseyowskiego antyreprezentacjonizmu.

Ad c) Określenie, czy „reprezentacje" desygnowane przez pojęcia wymienione w (b) funkcjonują w sposób na tyle zbliżony do sposobu funkcjonowania reprezentacji w sensie przednaukowym, aby można im było przypisać rolę reprezentowania

Ostatni krok metody Ramseya polega na ewaluacji, czy poszczególne wymienione w punkcie (b) kategorie „kandydatów" do miana reprezentacji rzeczywiście charakteryzują się odpowiednim profilem funkcjonalnym (to znaczy profilem sprawiającym, że jest uzasadnione przypisywanie im roli reprezentacji). Biorąc pod uwagę racje wymienione w punkcie (a), tego rodzaju ocena będzie w nieuniknion y sposób związana z koniecznością odwołania się do tego, jak rola

reprezentowania czegoś jest rozumiana zgodnie z naszymi przednaukowymi pojęciami reprezentacji. Aby sprostać wymogowi opisu zadań, przypisywanie danemu bytowi funkcji bycia reprezentacją w systemie poznawczym musi być naturalnie i intuicyjnie zrozumiałe z punktu widzenia tego, jak postrzega się rolę bycia reprezentacją w sensie przednaukowym. Jeśli warunek ten zostaje spełniony, mamy do czynienia z sytuacją, w której byt postulowany przez teorię jako reprezentacja rzeczywiście i w eksplanacyjnie wartościowy sposób pełni funkcję reprezentacji. Jeśli warunek ten nie zostanie spełniony, to powinniśmy orzec, że dana teoria odwołuje się do reprezentacji w sposób nieuprawniony; może ona (oraz powinna) obyć się bez posługiwania się terminologią reprezentacyjną. Wedle Ramseya (2007) tylko w tym pierwszym przypadku – kiedy wymóg opisu zadań zostaje spełniony – wyjaśnianie działania systemu poznawczego w kategoriach reprezentacyjnych daje nam rzeczywisty wgląd w jego działanie, realny eksplanacyjny zysk, którego nie uzyskalibyśmy, rezygnując z pojęcia reprezentacji.

Zauważmy, że metoda Ramseya zasadniczo różni się od procedury polegającej na wprowadzeniu technicznej definicji specyfikującej konieczne i wystarczające (funkcjonalne) warunki służenia jako reprezentacja w systemie poznawczym – tak jak robią to chociażby wspomniani Anderson i Rosenberg (2008). Jak już wskazywałem, to drugie, oparte na poszukiwaniu definicji podejście stwarza sytuację, w jakiej nietrudno wskazać przykłady struktur czy stanów, które spełniają podaną definicję, ale których rola funkcjonalna nie jest rozpoznawalnie reprezentacyjna. Wedle Ramseya (2007: 9–11) tego rodzaju propozycje zawodzą, ponieważ nie doceniają roli przednaukowego rodowodu pojęcia reprezentacji[19]. Autor ten wskazuje jednak jeszcze jeden powód takiego stanu rzeczy. Mianowicie każda próba wskazania koniecznych i wystarczających warunków pełnienia funkcji reprezentacji jest skazana na niepowodzenie. Dzieje się tak dlatego, że nasze przednaukowe pojęcie reprezentacji nie

[19] Dlaczego kontrprzykłady dla tych koncepcji tak łatwo i naturalnie rozpoznajemy właśnie jako kontrprzykłady? Dlatego, iż dana teoria każe nam je uznawać za przypadki reprezentacji – mimo faktu, że są one bardzo odległe od przednaukowego rozumienia reprezentacji.

jest mentalną definicją, lecz ma najprawdopodobniej naturę *prototypową*. Różnego rodzaju obiekty są klasyfikowane jako reprezentacje nie dlatego, że spełniają zestaw określonych koniecznych i wystarczających warunków, lecz na podstawie ich podobieństwa do prototypu czy prototypów (to znaczy na podstawie tego, że mają one odpowiednio wiele własności uznawanych za prototypowo reprezentacyjne). Dotyczy to także reprezentacji postulowanych przez kognitywistów. Dowolna wewnętrzna struktura systemu poznawczego czyni zadość wymogowi opisu zadań tylko wtedy, gdy naturalne i zrozumiałe jest postrzeganie roli przez nią pełnionej jako roli polegającej na reprezentowaniu czegoś. Ta ostatnia ocena jest dokonywana na podstawie funkcjonalnego podobieństwa zachodzącego między tą strukturą a prototypowymi przykładami reprezentacji.

Wykonana przez Ramseya w kroku (a) rekonstrukcja przednaukowych pojęć reprezentacji jest wartościowa dlatego, że pozwala ukierunkować opisywaną tu, realizowaną w kroku (c) procedurę. W kroku (a) zostały wymienione dwa ogólne rodzaje reprezentacji w przednaukowym znaczeniu tego terminu. Na tej podstawie możemy stwierdzić, że status pewnej wewnętrznej struktury jako reprezentacji w sensie naukowym będzie zależał od stopnia podobieństwa zachodzącego między rolami funkcjonalnymi pełnionymi przez tę strukturę a rolami funkcjonalnymi pełnionymi przez wskazane w punkcie (a) *przednaukowe prototypy* reprezentacji. Wzięcie tego założenia pod uwagę pozwala nadać pewien rygor krokowi (c). Metoda Ramseya nie sprowadza się do prostego przyjrzenia się „zakresowi zadań" pełnionych przez byt postulowany jako reprezentacja i intuicyjnej ocenie, czy realizowanie tych zadań rzeczywiście uzasadnia traktowanie czegoś jako reprezentację. Daje nam ona pojęcie, do czego powinny być (funkcjonalnie) podobne byty postulowane jako reprezentacje w kognitywistyce, aby zagwarantować sobie miano (rzeczywistych) reprezentacji.

Przedstawione wyżej twierdzenia generują jednak dość oczywisty problem (por. Ramsey 2007: 26–27). Kognitywiści zajmują się odkrywaniem i opisywaniem mechanizmów o charakterze neuro-

nalnym czy neuroobliczeniowym[20]. Jeśli tak, to przednaukowe pojęcia reprezentacji nie dadzą się w żaden oczywisty sposób zaaplikować w teoriach z zakresu kognitywistyki w niezmienionej formie. Dla przykładu mapy, wskazówki kompasów czy znaki języka naturalnego są reprezentacjami tylko o tyle, o ile służą jako reprezentacje *podmiotom intencjonalnym* czy też są jako reprezentacje przez te podmioty interpretowane. Kognitywiści nie mogą jednak – pod groźbą popełnienia błędu homunkularnego – postulować istnienia podmiotów intencjonalnych (czy jakichś ich funkcjonalnych odpowiedników) wewnątrz ośrodkowego układu nerwowego. Analogicznie, jeśli przyjrzymy się filozofii umysłu i problemom związanym z projektem naturalizacji intencjonalności, łatwo zauważymy, że nie jest oczywiste, jak niewywiedziona treść intencjonalna przypisywana postawom propozycjonalnym mogłaby być posiadana przez wewnętrzne stany naturalistycznie pojmowanych systemów poznawczych. W jaki sposób zatem działanie neuronalnych czy neuroobliczeniowych mechanizmów może w ogóle opierać się na strukturach (komponentach) funkcjonujących w sposób analogiczny do tego, jak funkcjonują reprezentacje w sensie przednaukowym? Czy metoda Ramseya nie nakłada zbyt restrykcyjnych warunków na bycie eksplanacyjnie wartościową reprezentacją w kognitywistyce?

Powyższe obserwacje nie stanowią jednak dla metody Ramseya takiego problemu, jak może się początkowo wydawać. Propozycja Ramseya nie jest przesadnie konserwatywna pojęciowo. Jak już wcześniej zaznaczyłem, wymóg opisu zadań nie wiąże się z wymogiem, aby reprezentacje w sensie naukowym funkcjonowały dokładnie tak, jak reprezentacje w sensie przednaukowym. Wystarczy, że funkcjonowanie tych pierwszych w nietrywialny i eksplanacyjnie wartościowy sposób przypomina funkcjonowanie tych drugich, na-

[20] Wyprzedzając rozstrzygnięcia pojęciowe czynione w dalszej części tej pracy, można powiedzieć, że na ile koncepcje tworzone w ramach nauk kognitywnych mają dostarczać wyjaśnień mechanistycznych, na tyle będą one formułowane na subosobowym poziomie opisywania i wyjaśniania systemu poznawczego. Tymczasem przednaukowe pojęcia reprezentacji wydają się mieć zastosowanie jedynie na poziomie osobowym, na którym mowa o intencjonalnych podmiotach i ich działaniach.

wet jeśli pod szeregiem innych względów zachodzą tu także różnice. Reprezentacje jako narzędzia eksplanacyjne kognitywistyki mogą okazać się swego rodzaju „mechanicznymi" wersjami map, wskazówek kompasów czy przekonań: być może są one w pewnym zakresie lub pod pewnymi względami wykorzystywane wewnątrz systemu poznawczego w sposób podobny do tego, jak są wykorzystywane (jakie role pełnią) mapy, kompasy czy przekonania. W ten właśnie sposób, jak to stwierdza sam Ramsey (2007: 31), sedno proponowanej przez niego metody polega na umożliwieniu pokierowania naszym myśleniem o reprezentacjach tak, byśmy znaleźli się między Scyllą koncepcji nakładających na reprezentacje zbyt mocne warunki (na przykład dlatego, że presuponują istnienie podmiotu intencjonalnego wewnątrz systemu poznawczego) a Charybdą koncepcji zbyt słabych, w których za reprezentacje są uznawane stany czy struktury, które nie pełnią roli reprezentacji w żadnym rozpoznawalnym sensie tego terminu.

Trzeba zwrócić uwagę na jeszcze jedną istotną kwestię. Metoda Ramseya jako taka wydaje się przydatna dla mechanicysty, ponieważ dostarcza wiarygodnej odpowiedzi na postawione wyżej pytanie 2: „Na jakiej podstawie możemy orzekać, że posiadanie określonych własności funkcjonalnych przez komponent mechanizmu gwarantuje mu status reprezentacji (nośnika reprezentacji)?". Mianowicie dostarcza ona kryterium odróżniania funkcji, które możemy w sposób uzasadniony uznać za polegające na reprezentowaniu czegoś, od takich, o których nie możemy tego (w sposób uzasadniony) orzec. Jednakże zauważmy, że krok (c) nie odwołuje się *explicite* do koncepcji mechanizmów reprezentacyjnych w zaproponowanym wcześniej rozumieniu[21]. Tymczasem w tej pracy centralne znaczenie ma dla mnie założenie, że kognitywistyczne wyjaśnienia za pomocą reprezentacji stanowią formę wyjaśnień mechanistycznych. Kiedy przyjmuje się taką perspektywę, należy uznać, że jeśli określony rodzaj reprezentacji spełnia wymóg opisu zadań, to powinniśmy być

[21] Ramsey (2007) mówi na ogół o reprezentacjach jako wewnętrznych „strukturach" lub „stanach". Nie zawsze jest oczywiste, jak tego rodzaju sposób wyrażania przeformułować na pojęcia zaczerpnięte z mechanicyzmu.

zdolni wyrazić naturę tych reprezentacji za pomocą narzędzi konceptualnych dostarczanych przez mechanicyzm. Mówiąc prościej, wyniki zastosowania metody Ramseya powinniśmy móc „przetłumaczyć" na język mechanistycznego modelu wyjaśnienia: język mechanizmów, komponentów mechanizmów, operacji wykonywanych przez komponenty oraz wewnętrznej organizacji takich aktywnych komponentów. Z punktu widzenia przyjmowanych tu przeze mnie założeń – rodzaje czy kategorie reprezentacji, w przypadku których wykonanie kroku (c) zakończyło się pozytywnie, powinniśmy uznać za *komponenty* wewnętrznych mechanizmów poznawczych. Jednocześnie komponenty takie powinny zawdzięczać swój status reprezentacji temu, jaką *rolę funkcjonalną* (operację) wykonują one *względem pozostałych komponentów* mechanizmu. Uzupełnijmy zatem metodę Ramseya następującym ostatnim krokiem:

> **d)** Jeśli dane pojęcie reprezentacji spełnia wymóg opisu zadań – czyli w jego przypadku wykonanie kroku (c) zakończyło się pozytywnym wynikiem – należy poddać je interpretacji za pomocą kategorii mechanistycznych, to znaczy: (1) przyjąć, że struktury egzemplifikujące to pojęcie są (mogą być) komponentami mechanizmów poznawczych; (2) wyrazić, dzięki jakiej roli funkcjonalnej pełnionej przez te komponenty w ramach mechanizmów poznawczych – komponenty te czynią zadość wymogowi opisu zadań (a tym samym pełnią funkcję reprezentacji).

W taki oto sposób uzupełniona metoda Ramseya okazuje się dla mechanicysty narzędziem służącym stworzeniu koncepcji mechanizmów reprezentacyjnych. Wykonanie kroku (d) pozwala bowiem wyrazić w klarowny i otwarty sposób, na czym polega specyfika funkcjonalnej „architektury" mechanizmów reprezentacyjnych[22]. Krok ten wykonam w kolejnym rozdziale, w kontekście rozważań

[22] Zauważmy, że wykonanie kroku (d) nie powinno być rozumiane jako wprowadzenie definicji reprezentacji. Na podstawie wspomnianych wcześniej uwag o prototypowej naturze przednaukowych pojęć reprezentacji powinniśmy uznać, że dostarczenie takiej definicji jest niewykonalne. Krok (d) należy więc raczej rozumieć jako procedurę polegającą na wskazaniu,

nad mechanizmami, których działanie opiera się na konsumowanych modelach. Pokażę tam najpierw, że S-reprezentacje – reprezentacje oparte na strukturalnym podobieństwie, czyli modele – spełniają wymóg opisu zadań: werdykt w kroku (c) jest pozytywny, co z kolei stanie się dla mnie podstawą dla sformułowania koncepcji mechanizmów reprezentacyjnych – krok (d). Zgodnie z tą koncepcją mechanizmy reprezentacyjne to takie, w skład których S-reprezentacje wchodzą jako ich aktywne komponenty.

3.3. Metoda Ramseya: przykłady zastosowań negatywnych

3.3.1. Krytyka reprezentacji receptorowych

Aby mieć całkowitą jasność co do natury metody Ramseya oraz roli, jaką może ona odgrywać w rozważaniach nad reprezentacjami, warto przyjrzeć się bliżej konkretnym przykładom jej praktycznego zastosowania. W następnym rozdziale metoda Ramseya zostanie zastosowana w celu ewaluacji eksplanacyjnego statusu S-reprezentacji, czyli reprezentacji opartych na podobieństwie strukturalnym. Będzie to przykład sytuacji, w której aplikacja metody Ramseya przynosi pozytywne rozstrzygnięcie: S-reprezentacje rzeczywiście służą w systemie poznawczym jako reprezentacje, w związku z czym mogą stanowić podstawę dla koncepcji mechanizmów reprezentacyjnych. Zauważmy jednak, że metoda oparta na wymogu opisu zadań jest szczególnie odkrywcza i interesująca ze względu na swój *rewizjonistyczny* potencjał. Stanowiłaby ona zapewne narzędzie dość nieciekawe, gdyby jej zastosowanie prowadziło do wniosku, że kognitywiści w sposób systematyczny i całościowy posługują się uprawnionymi, eksplanacyjnie wartościowymi pojęciami reprezentacji. Odkrywczość metody Ramseya polega na tym, że pozwala ona na znaczącą rewizję praktyki pojęciowej kognitywistów. Dzięki

w kategoriach mechanistycznych, jakie własności funkcjonalne komponentu mechanizmu gwarantują mu odpowiedni *stopień podobieństwa* do prototypowych reprezentacji.

jej zastosowaniu możliwa staje się identyfikacja i odrzucenie szeregu nieuprawnionych pojęć reprezentacji, które przy bliższej inspekcji okazują się eksplanacyjnie jałowe.

Negatywne zastosowania metody Ramseya są szczególnie interesujące także ze względu na główny cel teoretyczny tej pracy, jakim jest wypracowanie koncepcji wyjaśniania reprezentacyjnego. Pokazują one, że stworzenie funkcjonalnej koncepcji reprezentacji na potrzeby kognitywistyki to wyzwanie nietrywialne. Jeśli tak, to nietrywialnym zadaniem okazuje się też stworzenie teorii mechanizmów reprezentacyjnych. Nie da się tego zrobić przez dodanie drobnych aneksów do istniejących już wcześniej teorii reprezentacji mentalnych. Mechanizmy, których działanie opiera się na strukturach (komponentach) nazywanych powszechnie „reprezentacjami", bardzo często nie zasługują na miano mechanizmów reprezentacyjnych. Z tego powodu mechanistyczne wyjaśnienia zjawisk poznawczych za pomocą tego rodzaju mechanizmów nie są, wbrew powszechnie akceptowanym opiniom, wyjaśnieniami reprezentacyjnymi.

Przyjrzyjmy się zatem dwóm klarownym przykładom zastosowania metody Ramseya, które przynoszą negatywną diagnozę w odniesieniu do określonych, powszechnie stosowanych w kognitywistyce oraz filozofii pojęć reprezentacji. Jedno z nich to pojęcie reprezentacji jako receptorów, natomiast drugie to pojęcie reprezentacji ukrytych (niejawnych). Zacznijmy od tych pierwszych.

Zasadniczym elementem receptorowego pojęcia reprezentacji jest idea, że reprezentowanie opiera się na systematycznej współzmienności między nośnikiem reprezentacji a tym, co podlega reprezentowaniu. Nośnik reprezentuje pewne okoliczności (zdarzenie, obiekt czy stan rzeczy) dlatego, że systematycznie współzmienia się on z zajściem tych okoliczności; na przykład znajduje się on w określonym stanie kiedykolwiek zajdą określone okoliczności. Często przyjmuje się, że zachodzenie takiej współzmienności ma u swoich podstaw istnienie przyczynowych zależności między nośnikiem a tym, co reprezentowane[23]. Reprezentacje tego rodzaju są uznawane

[23] Warto nadmienić, że w filozofii umysłu istnieją pewne punkty sporne dotyczące tego, jak dokładnie charakteryzować przyczynowe współzmienności mające

za receptory czy detektory, dzięki którym organizmy potrafią selektywnie rozpoznawać zachodzenie określonych stanów rzeczy w swoim otoczeniu (pojawienie się pożywienia, drapieżnika, partnera reprodukcyjnego i tak dalej).

Kognitywistyczne pojęcie reprezentacji receptorowych jest wyraźnie zbliżone do wymienionego w tabeli 2 przednaukowego pojęcia reprezentacji *indeksowych*. Reprezentacje tego ostatniego rodzaju opierają się na tym, że pewne systematyczne zależności między zjawiskami są wykorzystywane przez podmiot w celu pozyskiwania informacji lub wiedzy o czymś. Na takiej zasadzie dym może stanowić dla kogoś reprezentację (wskaźnik) ognia, liczba słoi w pniu może informować kogoś o wieku określonego drzewa, a pozycja wskazówki kompasu – o położeniu północy magnetycznej. Receptorowe reprezentacje znajdowane *wewnątrz* systemów poznawczych także opierają się na systematycznej współzmienności, choć w ich przypadku nie może być mowy, pod groźbą popełnienia błędu homunkularnego, o byciu wykorzystywanym czy interpretowanym przez pełnoprawny podmiot intencjonalny. Reprezentacje tego rodzaju mają stanowić wewnętrzną i w pełni „mechaniczną" wersję reprezentacji wskaźnikowych.

Jak zauważa Ramsey (2007: 120–123), struktury postulowane jako reprezentacje oparte na współzmienności stanowią dość rozpowszechnione narzędzie eksplanacyjne w naukach kognitywnych. Klasycznym przykładem wykorzystania receptorowego pojęcia re-

stanowić podstawę reprezentacji. Dretske (1988) stawia wymóg, by relacja zależności przyczynowej między nośnikiem reprezentacji a tym, co reprezentowane, miała charakter prawa (by zależności między nośnikiem a przedmiotem reprezentacji stanowiły egzemplifikacje pewnego prawa). Autor ten rozwija zatem teorię reprezentacji, wykorzystując pojęcie zależności nomologiczno-przyczynowych. Z kolei Millikan (2002) kategorycznie odrzuca tego rodzaju nacisk na prawa, stwierdzając, że nie jest on dobrze umotywowany teoretycznie. „Reprezentujące" współzmienności występujące w przyrodzie (a w każdym razie współzmienności, które ta autorka uznaje za przykłady reprezentacji) nie egzemplifikują według Millikan żadnych praw. Rozstrzygnięcie tego sporu nie jest potrzebne ze względu na cele teoretyczne tej książki. Proponuję założyć ogólnie, że pojęcie receptorowe odwołuje się do współzmienności, a szczegółowe kwestie dotyczące natury tej współzmienności – pozostawić otwartymi.

prezentacji w kognitywistyce jest artykuł *Co żabie oko mówi żabiemu mózgowi?* (*What does the frog's eye tell the frog's brain?*) Jerome'a Lettvina i współpracowników (1959), dotyczący neuronów w korze wzrokowej żab, których aktywność jest systematycznie współzmienna z pojawianiem się owadów. Neurony te miałyby reprezentować obecność owada w bezpośrednim otoczeniu żaby. Podobny sposób myślenia o reprezentacjach jest także obecny w innych neuronaukowych badaniach nad „detektorowymi" własnościami neuronów znajdujących się w korach wzrokowych różnych gatunków organizmów. Powszechną praktykę stanowi przypisywanie nawet pojedynczym neuronom funkcji polegającej na reprezentowaniu (rozpoznawaniu czy kategoryzowaniu) bodźców (linii horyzontalnych, kolorów, a nawet twarzy ściśle określonych osób), których pojawienie się w polu wzrokowym wykazuje współzmienność ze stanami, w jakich znajduje się dana komórka (por.: Logothetis, Sheinberg 1996; Quiroga, Reddy et al. 2005). Innym rejonem kognitywistyki, w którym badacze korzystają z pojęcia reprezentacji jako receptorów, jest wedle Ramseya (2007) koneksjonizm. W tym przypadku zwraca się uwagę na fakt, że ekspozycja sieci koneksyjnej na podobne do siebie dane wejściowe (takie jak fonemy, słowa czy zdjęcia twarzy ludzkich) skutkuje podobnymi wzorami aktywności w pośredniczących warstwach neuronów. Tego rodzaju wzory aktywności miałyby stanowić – właśnie dzięki zachodzeniu wspomnianej systematycznej współzmienności – reprezentacje obiektów (kategorii obiektów), na które eksponowana jest sieć (por.: Sejnowski, Rosenberg 1987; Gorman, Sejnowski 1988).

Czy reprezentacje receptorowe zasługują na swoje miano, jeśli zastosować do nich wymóg opisu zadań? Pamiętajmy, że nie chodzi tu o problem, czy odwołanie do współzmienności pozwala na znaturalizowanie treści stanów mentalnych. Na obecne potrzeby teoretyczne nie ma większego znaczenia, czy reprezentacje oparte na współzmienności mogą być błędne, ani też to, czy gwarantują one odpowiednią „ziarnistość" treści (na przykład czy odwołanie do współzmienności pozwala na odróżnienie reprezentacji królików od reprezentacji nierozłączonych części królików). Wszystkie te zagadnienia wiążą się dość ściśle z problemem treści intencjonalnej, któ-

ry pozostawiam tu na boku. Metoda Ramseya ma rozwiązać inny problem, mianowicie czy *rola funkcjonalna* pełniona przez struktury postulowane jako reprezentacje receptorowe rzeczywiście polega na reprezentowaniu. Otóż wydaje się dość jasne, że koncepcja odwołująca się jedynie do współzmienności kompletnie nie radzi sobie z wymogiem opisu zadań. Po pierwsze, taka koncepcja właściwie nie mówi nic o rolach funkcjonalnych reprezentacji (nośnika reprezentacji) w ramach szerszego mechanizmu. Sama współzmienność może co prawda potencjalnie determinować treść reprezentacji, jednak to nie rozstrzyga kwestii tego, jaką rolę odgrywa w systemie nośnik tej treści (O'Brien, Opie 2004). Po drugie, gdybyśmy traktowali zachodzenie współzmienności jako warunek wystarczający do uczynienia czegoś reprezentacją, pojęcie to uległoby trywializacji. Mielibyśmy do czynienia z panreprezentacjonizmem: w kategoriach reprezentacyjnych należałoby interpretować każdą współzmienność zachodzącą we Wszechświecie (O'Brien, Opie 2004; Ramsey 2007: 124–127). Teza, że system poznawczy jest systemem reprezentacyjnym, okazałaby się nieciekawa, a w kategoriach reprezentacyjnych można by także wyjaśniać chociażby zachodzące w organizmie procesy trawienne czy termoregulacyjne. Pojęcie receptorowe w jego „czystej" wersji nie spełnia więc wymogu opisu zadań. Sposób rozumienia „reprezentacji" na podstawie tego pojęcia jest tak obcy przednaukowym sposobom rozumienia reprezentacji i ich ról funkcjonalnych (w tym reprezentacji indeksowych, czyli opartych na współzmienności[24]), że nie może być tu w ogóle mowy o reprezentowaniu.

Być może jednak istnieje sposób na zrehabilitowanie receptorowego pojęcia reprezentacji przez wprowadzenie do niego stosownych rozszerzeń. Należałoby je uzupełnić w taki sposób, by struktury systematycznie współzmieniające się z określonymi okolicznościami pełniły jednocześnie pewną rolę funkcjonalną wewnątrz systemu

[24] Pamiętajmy wszakże, że w przypadku (zewnętrznych) reprezentacji indeksowych, zachodzenie współzmienności nie jest warunkiem wystarczającym do bycia reprezentacją. Aby zyskać status reprezentacji, odpowiednie współzmienności (współzmieniające się z czymś struktury) muszą być wykorzystane przez pewien podmiot w funkcji reprezentacji.

czy mechanizmu. Ramsey (2007: 127–131) zauważa, że tego rodzaju teoretyczny „aneks" możemy odnaleźć – przynajmniej na pierwszy rzut oka – w teoriach, które łączą współzmiennościową koncepcję treści z elementami o charakterze *teleologicznym*. Autor ten powołuje się konkretnie na rozważania Freda Dretskego (1986, 1988) nad reprezentacjami błędnymi. Nie wdając się w niuanse złożonej propozycji Dretskego, wystarczy powiedzieć, iż opiera się ona na obserwacji, że sama współzmienność nie wystarcza do wyjaśnienia, jak powstają błędne reprezentacje. Aby te ostatnie były możliwe, zachodzenie systematycznej współzmienności między jakąś wewnętrzną strukturą (komponentem mechanizmu) a określonymi okolicznościami musi stanowić *funkcję* tej struktury w ramach pewnego systemu (mechanizmu). Inaczej mówiąc, struktura ta stanowi element pewnego szerszego kontekstu *po to*, aby systematycznie współzmieniała się z czymś innym[25].

Zilustrujmy powyższą ideę za pomocą przykładu wskazanego przez Dretskego (1986). Beztlenowe bakterie morskie potrafią kierować swój ruch w kierunku wody o sprzyjającej im, niskiej zawartości tlenu. Zdolność ta opiera się na działaniu organelli nazywanych „magnetosomami". Zawierają one kryształki magnetytu, których ułożenie jest systematycznie zależne od położenia północy magnetycznej względem bakterii. Między stanami, w jakich znajduje się magnetosom, a relatywnym położeniem północy magnetycznej zachodzi zatem systematyczna współzmienność. Współzmienność ta okazuje się ważna dlatego, że położenie północy jest zarazem skorelowane z miejscem występowania preferowanej przez bakterie, nisko utlenionej wody. Magnetosomy pozwalają organizmowi na „rozpoznanie", gdzie znajduje się preferowane środowisko oraz poruszanie się w jego kierunku. Organelle znajdują się w bakterii po to, by współzmieniać się z czymś innym – współzmienność ta stanowi ich funkcję. Przypomnijmy, że Dretskemu (1986, 1988) ten te-

[25] Odwołując się do technicznej terminologii zaproponowanej przez Dretskego (1988), zachodzenie współzmienności stanowi strukturyzującą przyczynę (*structuring cause*) istnienia współzmieniającego się komponentu w szerszym mechanizmie. Komponent ten znajduje się w mechanizmie dlatego, że współzmienia się z czymś innym.

leologiczny element służy przede wszystkim do sformułowania naturalistycznej koncepcji tego, jak możliwe są reprezentacje błędne. Mówiąc w skrócie, magnetosomy mogą reprezentować błędnie położenie wody o niskiej zawartości tlenu wtedy, gdy nie spełniają poprawnie swojej funkcji, to znaczy z jakichś powodów zaburzona jest ich współzmienność z położeniem północy magnetycznej (a jednocześnie z położeniem wody o niskiej zawartości tlenu). Jak jednak zauważa Ramsey (2007: 127–131), teleologiczny element koncepcji Dretskego może być także wykorzystany w innym celu, mianowicie pokazania, dlaczego właściwie mamy uznawać magnetosomy za pełniące funkcję reprezentacji. Dzięki odwołaniu się do teleologii możemy pokazać, iż struktury postulowane jako receptory czy detektory rzeczywiście odgrywają rolę reprezentacji. Otóż w świetle takiej propozycji magnetosomy są reprezentacjami, ponieważ są *wykorzystywane* w bakterii jako swoisty wewnętrzny kompas. To już nie sama współzmienność, lecz *funkcjonalna współzmienność* czyni z nich reprezentacje.

Podsumowując powyższe ustalenia, zgodnie z nowym, rozszerzonym sposobem rozumienia reprezentacji receptorowych – z reprezentacjami tego rodzaju mamy do czynienia, gdy spełnione są dwa warunki: (1) zachodzi współzmienność między stanami komponentu mechanizmu a określonymi okolicznościami, (2) dzięki zachodzeniu owej współzmienności komponent ten jest funkcjonalny dla szerszego mechanizmu. Czy tego typu modyfikacja pozwala jednak pojęciu reprezentacji receptorowych spełnić wymóg opisu zadań? Bez wątpienia taka nowa koncepcja reprezentacji receptorowych unika przynajmniej problemu panreprezentacjonizmu: wszakże nie wszystkie przypadki zachodzenia współzmienności w przyrodzie spełniają warunek (2). Wydaje się ponadto, że dodanie elementu teleologicznego pozwala też receptorowemu ujęciu reprezentacji oddać sprawiedliwość idei, zgodnie z którą reprezentacje oparte na współzmienności – aby w ogóle stać się nimi – powinny być wykorzystywane w roli reprezentacji.

Ramsey (2007: 131–140) przedstawia jednak mocny argument za tym, że tego rodzaju obserwacje nie pozwalają zmienić negatywnej diagnozy dotyczącej statusu reprezentacji receptorowych: nawet

funkcjonalne współzmienności nie stanowią reprezentacji. Strategia argumentacyjna wykorzystywana przez Ramseya na rzecz tej tezy jest prosta. Opiera się ona na wskazaniu przykładów, w których warunki (1) i (2) są spełnione, jednak w sposób oczywisty i niekontrowersyjny nie mamy do czynienia z reprezentacjami (strukturami pełniącymi funkcję reprezentacji). Naciśnięcie spustu w pistolecie systematycznie uruchamia iglicę, a ta, uderzając w spłonkę, doprowadza do wystrzału. Funkcją iglicy jest inicjowanie wystrzału zawsze wtedy, gdy zostanie naciśnięty spust. Iglica pistoletu spełnia obydwa wymienione warunki – jej *funkcją* jest *współzmienianie* się ze stanami, w jakich znajduje się spust pistoletu. Co jednak jasne, nie pełni ona w pistolecie roli reprezentacji. Przykłady tego rodzaju można mnożyć. Układy termoregulacyjne w organizmach, ekosystemy, telewizory, koparki, krany, latarki, silniki[26] – działanie wszystkich tych oraz wielu innych systemów fizycznych opiera się na komponentach, których funkcja polega na współzmienianiu się z zajściem określonych okoliczności. Dlaczego mielibyśmy jednak wyjaśniać działanie tych systemów za pomocą reprezentacji? Jak stwierdza Ramsey, funkcjonalne współzmienności w realnych systemach fizycznych to raczej przypadki zwykłego pośredniczenia przyczynowego (*causal mediation*), w których okoliczność A wpływa na stan pewnej struktury B, co z kolei powoduje zajście pewnej okoliczności C. B jest w systemie czy mechanizmie po to, by „pośredniczyć"

[26] Warto na marginesie przypomnieć, że funkcjonowanie wspomnianego w rozdziale 1 silnika Watta opierało się na wewnętrznym regulatorze, który sprawiał, że prędkość ruchu tłoka w cylindrze oraz prędkość ruchu koła zamachowego wzajemnie się dostosowywały. Regulacja ta była możliwa dzięki systematycznej współzmienności zachodzącej między kątem nachylenia ramion regulatora a prędkością tłoka. Jak pamiętamy, właśnie ten fakt stanowi wedle Bechtela (1998) podstawę, aby sądzić, że silnik Watta to system reprezentacyjny. Regulator miałby reprezentować prędkość silnika „dla" koła zamachowego. Przedstawione tu argumenty przeciwko receptorowemu pojęciu reprezentacji w naturalny sposób uderzają w propozycję Bechtela oraz sprzyjają stanowisku, jakie w sporze o naturę silnika Watta zajmuje van Gelder. Regulator, w jaki jest wyposażony silnik Watta, to po prostu przyczynowy pośrednik/mediator między dwoma elementami silnika (tłokiem i kołem zamachowym). Nie ma dobrych podstaw, aby rozumieć jego rolę funkcjonalną jako rolę polegającą na reprezentowaniu.

przyczynowo między A i C, jednak nie ma żadnych dobrych powodów, żeby mówić, iż *reprezentuje* ono A *dla* C. Terminologia reprezentacyjna okazuje się tu zbędna i nie daje żadnego eksplanacyjnego zysku. Zauważmy jednak, że zasada działania rzekomych reprezentacji receptorowych w systemach biologicznych jest dokładnie taka sama[27]. Magnetosomy w bakteriach morskich spełniają warunki (1) i (2) oraz grają określoną rolę eksplanacyjną właśnie dlatego, że je spełniają. Odgrywana przez nie w systemie poznawczym rola funkcjonalna jest taka sama, co rola przyczynowych pośredników w systemach czy mechanizmach, których nigdy nie „podejrzewalibyśmy" o posługiwanie się reprezentacjami. Możemy zrozumieć zachowania bakterii bez przypisywania ich komponentom roli reprezentacji, a pozostając jedynie na poziomie postrzegania ich w kategoriach „pośredników" przyczynowych. Nie istnieją jakieś dobre racje za stosowaniem tu wyjaśnień reprezentacyjnych, dokładnie z tych samych powodów, dla których nie uznajemy za uzasadnione wyjaśniać w kategoriach reprezentacyjnych działania pistoletów.

Przytoczona krytyka pojęcia receptorowego pokazuje, że spełnienie warunków (1) i (2) nie wystarcza do bycia reprezentacją. Nie każdy przypadek, w którym regularną współzmienność wykorzystuje się w celu pełnienia jakiejś funkcji, jest zarazem przypadkiem sytuacji, w której funkcja ta polega na reprezentowaniu czegoś. Być może współzmienność mogłaby być wykorzystana w celu reprezentowania czegoś, ale sam fakt, że jest wykorzystywana w jakiejś funkcji, nie przesądza, iż jest to funkcja bycia reprezentacją[28]. We

[27] Twierdzenie to dotyczy jednak także systemów sztucznych (artefaktów), które bywają wskazywane jako przykłady systemów posługujących się reprezentacjami receptorowymi. W świetle krytyki Ramseya termostaty – stanowiące wedle Dretskego (1988) przykład prostych systemów korzystających z reprezentacji receptorowych – nie są systemami reprezentacyjnymi w wyższym stopniu niż pistolety.

[28] Można zilustrować tę myśl jeszcze innym, obrazowym przykładem zaczerpniętym z pracy Ramseya (2007: 135–136). Wyobraźmy sobie osobę, która zasadza drzewo w swoim ogródku. Robi to po to, by określona część ogródka była zacieniona o określonej porze dnia. Osoba ta wybiera miejsce zasadzenia drzewa, wykorzystując systematyczną zależność między pozycją słońca o danej porze dnia a długością cienia, jaki będzie zapewniać drzewo o tej porze. Mamy tu do

wszystkich wymienionych wyżej przykładach – włączając w to magnetosomy – funkcja ta polega nie na reprezentowaniu, lecz na „mediacji" przyczynowej.

Raz jeszcze można by jednak stwierdzić, że da się odeprzeć krytykę Ramseya przez wprowadzenie kolejnego uzupełnienia do naszego sposobu rozumienia reprezentacji receptorowych. Być może odpowiednie jej rozszerzenie pozwoliłoby na odróżnienie przypadków, w których struktura spełniająca warunki (1) i (2) pełni funkcję reprezentacji, od przypadków, w których jest wykorzystywana w jakiejś innej, niereprezentacyjnej funkcji (na przykład pośrednika przyczynowego). Wydaje się na pierwszy rzut oka, że istnieje możliwość wykonania ruchu tego rodzaju. Przedstawiona wyżej rekonstrukcja stanowiska Dretskego w ogóle nie odwoływała się bowiem do bardzo ważnej dla tego autora kategorii *informacji*. Być może wartościowe pojęcie reprezentacji receptorowych powinno nakładać na reprezentacje nie dwa, lecz trzy warunki: (1) zachodzenie współzmienności między komponentem mechanizmu a określonymi okolicznościami; (2) kodowanie (dzięki zachodzeniu wspomnianej współzmienności) określonej informacji przez ten komponent; (3) fakt, że kodowanie informacji stanowi funkcję tego komponentu w szerszym mechanizmie (por. Ramsey 2007: 132–134). Nie sama współzmienność, ale niesiona dzięki niej informacja miałaby zatem istotne znaczenie dla funkcjonowania danego komponentu mechanizmu[29]. To zaś gwarantowałoby mu status reprezentacji.

Ramsey (2007: 134–140) jednakże argumentuje, że także powyższe rozszerzenie receptorowego rozumienia reprezentacji nie

czynienia z wykorzystaniem współzmienności między pozycją słońca na niebie a długością cienia rzucanego przez drzewo. Co jednak jasne, w omawianej sytuacji owa współzmienność jest wykorzystywana nie w celu reprezentowania czegoś, lecz zapewnienia cienia określonemu miejscu w ogródku. Współzmienność ta mogłaby zostać wykorzystana jako reprezentacja, gdyby ktoś chciał dowiedzieć się, gdzie na niebie znajduje się słońce na podstawie długości cienia rzucanego przez drzewo. Jednak nie jest tak w tym przypadku. Warunki (1) i (2) są spełnione, ale nie może być mowy o reprezentowaniu.

[29] Dla przykładu, to nie sama współzmienność, ale informacja o tym, gdzie znajduje się woda o określonych własnościach, miałaby stanowić o funkcji pełnionej przez magnetosomy w bakteriach morskich.

jest pomocne. Zauważa on, iż wszystkie podawane przez Dretskego przykłady wykorzystania informacji przez sztuczne lub naturalne systemy są *de facto* przykładami wykorzystania systematycznych współzmienności między pewną wewnętrzną strukturą (komponentem mechanizmu) systemu a okolicznościami zewnętrznymi. W teorii Dretskego informacja jest całkowicie determinowana przez zachodzenie pewnej współzmienności; inaczej, niesienie informacji okazuje się tu tym samym, co współzmienianie się z czymś (względnie okazuje się ono całkowicie epifenomenalne – por. Ramsey 2007: 134–140). Z perspektywy tej teorii nie są możliwe przypadki, w których funkcją danego komponentu w pewnym szerszym mechanizmie jest współzmienianie się z czymś innym, ale jednocześnie funkcja ta nie polega na niesieniu informacji o tym, z czym ten komponent się współzmienia. Zawsze gdy mamy do czynienia z funkcjonującą współzmiennością, mamy też do czynienia z wykorzystaniem informacji, a zatem (według teorii) z reprezentowaniem. Biorąc to pod uwagę, okazuje się, że odwołanie się do informacji nie ratuje receptorowego pojęcia reprezentacji. Cała rola eksplanacyjna nadal jest „wykonywana" przez funkcjonalne współzmienności. Tym samym można tu zastosować wszystkie przedstawione wcześniej kontrprzykłady oraz wyciągnąć na ich podstawie identyczny wniosek.

Powyższe fakty nie oznaczają jeszcze, że przeprowadzenie linii demarkacyjnej między współzmiennościami pełniącymi funkcję reprezentacji a takimi, które jej nie pełnią, jest niemożliwe. Dotychczasowy wywód ma raczej skłaniać do wniosku, że linii tej nie da się przeprowadzić w taki sposób, by kowariancyjne/receptorowe pojęcie reprezentacji mogło być użyteczne dla celów eksplanacyjnych kognitywistyki. Jedyne niekontrowersyjne przykłady sytuacji, w których systematyczne współzmienności pełnią funkcję reprezentacji, to sytuacje, w jakich rola ta polega na informowaniu o czymś podmiotu intencjonalnego czy też na byciu wykorzystywanym przez podmiot w celu wyciągnięcia wniosków na temat pewnego obiektu (zdarzenia, stanu rzeczy i tak dalej). Wszelkie próby „pozbycia się" podmiotu intencjonalnego z pojęcia reprezentacji opartego na współzmienności sprawiają, że struktury czy komponenty współzmieniające się z czymś innym przestają pełnić rolę reprezentacji,

a okazują się czymś mniej teoretycznie czy eksplanacyjnie ciekawym – zwykłymi „pośrednikami" przyczynowymi. Biorąc zatem pod uwagę, że na poziomie opisu wewnętrznych mechanizmów poznawczych nie możemy postulować istnienia interpretujących podmiotów, okazuje się, iż pojęcie reprezentacji receptorowych nie spełnia wymogu opisu zadań. Bliższe spojrzenie na funkcję pełnioną przez receptory ukazuje, że rola ta nie może być w zrozumiały i intuicyjny sposób opisana jako „reprezentowanie czegoś". Zastosowanie metody Ramseya przynosi tu zatem negatywną diagnozę, w związku z czym należy uznać, że koncepcji mechanizmów reprezentacyjnych nie sposób oprzeć na idei reprezentacji receptorowych.

3.3.2. Krytyka reprezentacji ukrytych

Reprezentacje ukryte (niejawne) w rekonstrukcji Ramseya (2007: 151–167) nie posiadają w systemie poznawczym odrębnych, możliwych do zlokalizowania wewnętrznych nośników. Operowanie pojęciem reprezentacji ukrytych przez kognitywistów polega na przypisywaniu systemowi stanów mających treść, mimo że nie istnieje żadna wewnętrzna, możliwa do wyodrębnienia struktura (komponent mechanizmu), o której moglibyśmy powiedzieć, że stanowi ona nośnik tej właśnie, a nie innej treści. Reprezentacje tego rodzaju są ukryte (niejawne, implicytne), ponieważ nie są otwarcie (jawnie, eksplicytnie) kodowane przez wyodrębnione komponenty czy elementy systemu. Wewnętrzną, funkcjonalną architekturę systemu[30] rozpatruje się w kategoriach reprezentacyjnych nie dlatego, że możemy w niej wyodrębnić stany czy struktury pełniące funkcję reprezentacji, ale dlatego, że architektura ta stoi u podstaw określonych *zdolności* systemu. Mówiąc inaczej, niektóre zdolności systemu – rozumiane jako jego dyspozycje do określonych zachowań w określonych okolicznościach – mogą być wyjaśniane przez przypisywanie systemowi dysponowania reprezentacjami ukrytymi czy niejawnymi. Te ostatnie nie mają co prawda odrębnych nośników, jednak mo-

[30] Architekturę tę możemy uznać za ogół wewnętrznych mechanizmów, które stoją u podstaw zdolności poznawczych posiadanych przez system.

żemy o nich powiedzieć, że są „wcielone" w całość wewnętrznej architektury badanego systemu – atrybucja ukrytych reprezentacji jest oparta na fakcie, że architektura ta całościowo sprawia, iż system ma taki, a nie inny zestaw własności dyspozycyjnych (zdolności).

Powyższą, ogólną charakterystykę warto zilustrować przykładem praktycznego zastosowania pojęcia reprezentacji ukrytych w kognitywistyce. Choć pojęcie to było wykorzystywane w ramach klasycznego, symboliczno-obliczeniowego podejścia do modelowania systemu poznawczego[31], to według Ramseya (2007: 156–160) rozpowszechniło się ono w kontekście modelowania koneksjonistycznego (por.: Churchland 1989; Clark 1993; Rogers, McClelland 2004). Na czym polega wyjaśnianie działania sieci koneksyjnych przez postulowanie ukrytych reprezentacji? Za zachowanie dowolnej sieci koneksyjnej odpowiada układ wag[32] połączeń między sztucznymi neuronami znajdującymi się w pośredniczących warstwach tej sieci. Sieć ma określoną „zdolność" wtedy, gdy układ wag umożliwia jej posiadanie określonych (pożądanych przez projektanta) własności dyspozycyjnych. Otóż działanie sieci może być wyjaśniane w kategoriach reprezentacji ukrytych właśnie dlatego, że ma ona takie, a nie inne własności dyspozycyjne. Jeśli wewnętrzna struktura sieci pozwala jej, powiedzmy, na dyskryminowanie określonych bodźców (selektywne reagowanie na nie), to można powiedzieć, że sieć wykorzystuje ukryte reprezentacje tych bodźców. Reprezentacje te miałyby być całościowo „wcielone" w wewnętrzną strukturę sieci – to znaczy w układ wag połączeń między neuronami znajdującymi się w warstwach pośredniczących – i w taki właśnie sposób odpowiadać za jej działanie. Mówiąc technicznie, ukryte reprezentacje postulowa-

[31] W ramach tego podejścia posługiwanie się pojęciem reprezentacji ukrytych polegało na przypisywaniu systemowi poznawczemu stanów reprezentacyjnych posiadających treści, których jednak nie można przypisać żadnym symbolom czy ciągom symboli (strukturom danych) uczestniczącym w procesach obliczeniowych dokonujących się w systemie. W domyśle system miałby reprezentować pewne treści dzięki temu, że posiada określony profil behawioralny, a nie ze względu na treści *explicite* kodowane przez wewnętrzne symbole (por.: Dennett 1978; Clapin 2002).

[32] Wagi te określają eskcytacyjne i inhibitacyjne relacje między połączonymi neuronami sieci.

ne przez badaczy zaangażowanych w modelowanie koneksjonistyczne miałyby cechować się tym, że są one (a) nie tyle zlokalizowane, co raczej *rozproszone* (*distributed*) po całej sieci oraz (b) *nałożone* (*superimposed*), w tym sensie, że jeden układ wag połączeń służy do reprezentowania wielu różnych treści (Ramsey 2007: 156–167).

Weźmy dla przykładu pod uwagę prostą sieć koneksyjną, która jest eksponowana na zdjęcia ludzkich twarzy (por. Golomb, Sejnowski 1995; por. także rozpoznającą twarze sieć Cottrela, której działanie zostanie omówione w sekcji 4.3). Cel projektanta stanowi „wyuczenie" sieci zdolności do selektywnego reagowania na dowolne zdjęcie w zależności od płci przedstawionej na nim osoby. Jeśli w wyniku treningu zostaje ustalony układ wag, który dysponuje sieć do innego reagowania na twarze męskie niż żeńskie, sieci tej może zostać przypisane posiadanie ukrytych, kategorialnych *reprezentacji* twarzy męskich i żeńskich. Sieć miałaby się posługiwać reprezentacjami tego rodzaju pomimo faktu, że nie ma w niej żadnych możliwych do wyodrębnienia struktur (pojedynczych neuronów lub ich grup), którym można by przypisać określoną treść („twarz męska" lub „twarz żeńska"). To cały układ wag w warstwach pośredniczących miałby reprezentować te kategorie, w związku z czym mielibyśmy do czynienia z reprezentacjami rozproszonymi. Jednocześnie ten sam układ wag byłby zaangażowany w reprezentowanie zarówno twarzy męskich, jak i żeńskich, a więc reprezentacje, o których mowa, miałyby być także nałożone.

Czy reprezentacje ukryte czynią zadość wymogowi opisu zadań? Szybko okazuje się, że takie ujęcie jest zdecydowanie zbyt liberalne. Po pierwsze, zapytajmy, co każe nam sądzić, że posiadanie pewnych dyspozycji przez system czy obiekt fizyczny wystarcza do tego, aby wyjaśniać działanie tego systemu za pomocą reprezentacji (Ramsey 2007: 167–168)? Struktura molekularna szkła odpowiada za dyspozycję szklanki do stłuczenia się po uderzeniu w podłogę, jednak nie ma dobrych racji, by uznać, że struktura ta cokolwiek reprezentuje – i to w jakikolwiek, jawny czy ukryty sposób. Czy kostka cukru posiada wewnętrzne reprezentacje, które odpowiadają za fakt, że ma ona dyspozycję do rozpuszczenia się w wodzie? Nie istnieje takie przednaukowe rozumienie reprezenta-

cji, które skłaniałoby nas do myślenia, że u podstaw dowolnej dyspozycji stoją (ukryte) reprezentacje.

Po drugie, nawet zawężenie rozważań do dyspozycji posiadanych przez *stricte poznawcze* systemy nie pomaga pojęciu reprezentacji ukrytych. Można trywialnie zauważyć, że systemy poznawcze zawsze mają pewną wewnętrzną architekturę funkcjonalną. Mechanizmy składające się na nią odpowiadają za zdolności systemu, to znaczy za dyspozycje tego systemu, jako całości, do określonego działania. Jeśli posługujemy się pojęciem reprezentacji ukrytych, okazuje się, że każdy możliwy zestaw mechanizmów składających się na architekturę systemu będzie można rozpatrywać w kategoriach reprezentacyjnych. Skoro mechanizmy te są odpowiedzialne za ogólny dyspozycyjny „profil" systemu, można im przypisywać posługiwanie się ukrytymi reprezentacjami, nawet jeśli nie istnieją żadne możliwe do wyodrębnienia komponenty pełniące w nich role (nośników) reprezentacji. O naturze mechanizmu jako reprezentacyjnego będzie decydowało nie to, jak wyjaśnia on zdolności poznawcze – to, czy jakieś jego komponenty są zaangażowane w reprezentowanie, czy nie – ale sam fakt, *że* je on wyjaśnia. Jest to niepożądana konsekwencja. W takim wypadku reprezentacjonizm staje się stanowiskiem trywialnym, a antyreprezentacjonizm – niemożliwym do obrony. Ten pierwszy byłby kompatybilny z każdą możliwą koncepcją wewnętrznej architektury systemu poznawczego[33]. Jednocześnie niemożliwa byłaby obrona stanowiska antyreprezentacjonistycznego, bo każdy mechanizm – nawet taki, który nie ma żadnego komponentu zaangażowanego funkcjonalnie w reprezentowanie – mógłby zostać uznany za posługujący się reprezentacjami ukrytymi. W ostatecznym rozrachunku pojęcie reprezentacji ukrytych nie stwierdza o wewnętrznej architekturze systemu poznawczego nic więcej, niż tylko oczywisty fakt, że architektura ta stoi u podstaw dyspozycji posiadanych przez ten system jako całość (Ramsey 2007: 167). Jak się jednak wydaje, choć reprezentacje (o ile istnieją)

[33] Aby na przykład uznać, że system działa na podstawie reprezentacji okoliczności *X*, wystarczyłoby stwierdzić, iż jest on tak wewnętrznie ustrukturyzowany, by mieć odpowiednie własności dyspozycyjne względem *X* (dyspozycję do dyskryminacji *X*, do adaptywnego reagowania na *X* i tak dalej).

odpowiadają za określone własności dyspozycyjne (zdolności) systemu poznawczego, to jednak tym, *co czyni je reprezentacjami*, jest pewien specyficzny *sposób*, w jaki odpowiadają one za te własności – pewna rola funkcjonalna, jaką one pełnią wewnątrz systemu.

Czy istnieje dla zwolennika reprezentacji ukrytych jakiś sposób na to, aby zneutralizować powyższe zarzuty? Ramsey (2007: 152––156) twierdzi, że zachodzi związek między kognitywistycznym pojęciem reprezentacji ukrytych a przednaukowym, osadzonym w naszej psychologii potocznej pojęciem *wiedzy proceduralnej* czy też „wiedzy, jak" (odróżnionej od „wiedzy, że"). Dana osoba może wiedzieć, jak wykonywać pewną czynność praktyczną (pływać, jeździć na rowerze, szczotkować zęby i tak dalej), nie posiadając jednocześnie przekonań *explicite* dotyczących reguł czy zasad wykonywania tej czynności. Tym samym niektóre zdolności praktyczne posiadane przez ludzi są oparte na specyficznym rodzaju praktycznej wiedzy, która nie jest „przechowywana" w formie przekonań czy innych stanów intencjonalnych (czyli w formie „wiedzy, że"). Według Ramseya idea proceduralnej czy praktycznej wiedzy stoi u podstaw kognitywistycznego pojęcia reprezentacji ukrytych. Posiadanie określonych własności dyspozycyjnych (zdolności) przez system poznawczy byłoby w takiej perspektywie jednoznaczne z dysponowaniem przez ten system pewną wiedzą, konkretnie wiedzą o charakterze proceduralnym, która nie musi być zakodowana w określonych, wewnętrznych nośnikach. Jednakże fakt, że mamy tu do czynienia z wiedzą miałby sprawiać, iż także przypisywanie systemowi *reprezentacji* staje się uzasadnione. Bowiem wiedza bez jakiejś formy reprezentacji wydaje się niemożliwa. Argument za koniecznością postulowania reprezentacji ukrytych mógłby zatem wyglądać następująco (za: Ramsey 2007: 168–169)[34]:

1. Wewnętrzna, funkcjonalna architektura systemu poznawczego jest przyczynowo odpowiedzialna za zdolności poznawcze

[34] Ramsey konstruuje ten argument na podstawie rozważań Johna Haugelanda (1998) nad naturą reprezentacji ukrytych. Argument ten został tu dla celów poglądowych nieco uproszczony w stosunku do tego, jaką formę przyjmuje on w pracy Ramseya.

(dyspozycje do określonego działania) posiadane przez ten system.

2. Stąd w funkcjonalną architekturę systemu poznawczego jest „wcielona" (*embodied*) pewnego rodzaju wiedza (mianowicie wiedza proceduralna czy też „wiedza, jak").

3. Wiedza jest niemożliwa bez jakiegoś rodzaju reprezentacji.

4. Stąd w funkcjonalną architekturę systemu poznawczego są „wcielone" pewnego rodzaju reprezentacje.

5. Na funkcjonalną architekturę systemu nie składają się jawne reprezentacje.

6. Stąd architektura funkcjonalna systemu poznawczego wykorzystuje reprezentacje ukryte.

Ciężar powyższego argumentu spoczywa na twierdzeniu, że posiadanie określonych zdolności przez system poznawczy jest równoznaczne z dysonowaniem przezeń pewnym rodzajem wiedzy. Klucz do odparcia tego rozumowania stanowi zwrócenie uwagi na fakt, że termin „wiedza" został w nim użyty niejednoznacznie (Ramsey 2007: 169). Możemy go bowiem w kontekście powyższego rozumowania interpretować na dwa sposoby. Zgodnie z pierwszą, słabą interpretacją posiadanie wiedzy (dokładniej „wiedzy, jak") przez system poznawczy to nic więcej, niż posiadanie przezeń określonych zdolności (własności dyspozycyjnych). Zgodnie z drugą, mocną interpretacją posiadanie zdolności określanych jako „wiedza" wymaga od systemu posługiwania się wewnętrznymi reprezentacjami[35].

Zauważmy jednak, że przy dowolnej z tych dwóch interpretacji „wiedzy", przedstawiony wyżej argument okazuje się niekonkluzywny (por. Ramsey 2007: 169–173). Jeśli przyjmiemy interpretację słabą, przesłanka (3) będzie fałszywa. Posiadanie pewnych dyspozycji przez określony system fizyczny – co miałoby być równoznaczne z posiadaniem przez niego pewnego rodzaju wiedzy – w żaden sposób nie implikuje czy gwarantuje, że system ten (nawet jeśli mowa

[35] System w tym przypadku musi reprezentować to, jak wykonać pewną czynność czy zadanie. Inaczej mówiąc, działania, które stanowią manifestację posiadanej przez system „wiedzy, jak", muszą być w tym przypadku przewodzone za pomocą reprezentacji specyfikujących określone reguły praktyczne.

o systemie poznawczym) posługuje się reprezentacjami. Racje za taką tezą zostały już przedstawione powyżej. Jeśli mamy uczciwie traktować antyreprezentacjonizm, nie możemy *a priori* zakładać, że pewnych zjawisk nie sposób wyjaśnić bez reprezentacji. Z kolei jeśli nadamy terminowi „wiedza" mocną interpretację, to okazuje się co prawda, że przesłanka (3) nie jest problematyczna, jednakże przejście z (1) do (2) staje się nieuprawnione. Nie możemy wszakże *presuponować*, że posiadanie zdolności poznawczych przez pewien system jest jednoznaczne z dysponowaniem przezeń wiedzą w mocnym sensie, a zatem (*ex definitione*) – z posługiwaniem się reprezentacjami. Innymi słowy, przy mocnej interpretacji „wiedzy" przejście z (1) do (2) wymaga założenia tego, co miało zostać dowiedzione. Skoro antyreprezentacjonizm to realna opcja, nie możemy *a priori* wykluczać możliwości, że zdolności posiadane przez systemy poznawcze nie opierają się na działaniu mechanizmów reprezentacyjnych.

Jak się zatem okazuje, omawiany argument opiera się na ekwiwokacji i w ostatecznym rozrachunku nie pozwala na skuteczną obronę eksplanacyjnego statusu reprezentacji ukrytych. Obydwie przedstawione wcześniej racje przeciwko reprezentacjom ukrytym i ich eksplanacyjnemu statusowi pozostają w mocy. Pojęcie reprezentacji ukrytych okazuje się zatem nieuprawnione. Reprezentacje tego rodzaju nie zdają egzaminu: pełnione przez nie „zadania" nie polegają na reprezentowaniu czegokolwiek. Reprezentacje ukryte, podobnie jak receptorowe, nie spełniają wymogu opisu zadań.

Przedstawiona w tym podrozdziale rekonstrukcja siłą rzeczy pomija pewne niuanse argumentacji Ramseya przeciwko reprezentacjom receptorowym i ukrytym. Zasadnicze elementy rozumowania tego autora są w niej jednak zawarte. Aby określić status eksplanacyjny reprezentacji postulowanych przez daną teorię, należy najpierw skrupulatnie przyjrzeć się rolom pełnionym przez nie w szerszym systemie czy mechanizmie. Jeżeli role te nie polegają w intuicyjny i zrozumiały sposób na reprezentowaniu czegoś, to nie zasługują na miano reprezentowania. Zgodnie z naszymi przednaukowymi sposobami rozumienia reprezentacji do pełnienia tej roli nie wystarcza ani to, że dana struktura jest „pośrednikiem" przyczynowym, ani to,

że dana struktura stoi u podstaw dyspozycji systemu poznawczego[36]. Nie wykluczam tu możliwości, że pojęcia reprezentacji receptorowych lub ukrytych są do „uratowania". Być może da się je uzupełnić i dookreślić tak, by w ich kontekście można było w sposób rzetelny i uzasadniony mówić o reprezentowaniu. Jednak przyjmuję tu jednocześnie, że Ramseya rekonstrukcja tych pojęć reprezentacji jest uczciwa, a ich krytyka – trafna. Skoro tak, to pojęcia reprezentacji receptorowych oraz ukrytych wymagają istotnych modyfikacji lub – jeśli te nie pomogą – całkowitego odrzucenia. W każdym razie ciężar argumentacji spoczywa na zwolennikach idei, że systemy poznawcze posługują się jednym z tych typów reprezentacji. Na razie możemy przyjąć, że „reprezentacje" receptorowe i ukryte są nimi tylko z nazwy, a nie ze względu na role funkcjonalne, jakie rzeczywiście pełnią. Pozwala to skonstatować, że jeśli poszukuje się koncepcji mechanizmów reprezentacyjnych czyniącej zadość wymogowi opisu zadań, należy skierować swoje poszukiwania w inną stronę.

[36] Warto na marginesie zauważyć, że ustalenia tego podrozdziału uderzają także w opisaną wcześniej koncepcję reprezentacji jako przewodników działań Andersona i Rosenberga (2008). Zarówno „reprezentacje" receptorowe, jak i „reprezentacje" ukryte przewodzą działaniami systemu poznawczego. (Różnią się one jednak tym, czy w systemie występuje osobna, zlokalizowana struktura stanowiąca nośnik reprezentacji). Skoro jednak zarazem struktury tego rodzaju nie pełnią tak naprawdę roli reprezentacji – nie przewodzą działaniami systemu jako reprezentacje – to przewodzenie działaniami nie wystarcza do bycia reprezentacją.

ROZDZIAŁ 4

Koncepcja mechanizmów reprezentacyjnych

4.1. Od zewnętrznych reprezentacji ikonicznych do wewnętrznych S-reprezentacji

4.1.1. Reprezentacje w kognitywistyce a Peirce'owska triada

Dotychczasowy wywód ma stanowić teoretyczny fundament, na którym chcę zbudować pozytywną koncepcję wyjaśniania reprezentacyjnego w kognitywistyce. Otóż teza główna tej książki, stanowiąca zarazem główną tezę obecnego rozdziału, brzmi:

> Mechanizmy reprezentacyjne to mechanizmy, które są wyposażone w konsumowane modele. Reprezentacyjne wyjaśnienie pewnego zjawiska (zdolności) to wyjaśnienie go za pomocą mechanizmu wyposażonego w konsumowany model.

Nieco bardziej technicznie: mechanizmy reprezentacyjne to takie mechanizmy, w skład których wchodzą komponenty stanowiące *konsumowane modele* pewnej (reprezentowanej) domeny. W związku z tym *wyjaśnienie reprezentacyjne* – rozumiane jako wyjaśnienie zjawisk za pomocą mechanizmów reprezentacyjnych – odwołuje się do mechanizmu wyposażonego w konsumowany model (posiadający komponent pełniący funkcję modelu). Aby uzasadnić tę tezę, chcę pokazać, że konsumowane modele z powodzeniem przechodzą test odwołujący się do opisanego w poprzednim rozdziale wymogu opisu zadań. Oznacza to, że konsumowane modele pełnią w mechanizmach rolę funkcjonalną, która – w uzasadniony, eksplanacyjnie wartościowy sposób – polega na reprezentowaniu czegoś. Celem tego rozdziału będzie obrona oraz dostarczenie jasnego i precyzyj-

nego sformułowania koncepcji mechanizmów reprezentacyjnych jako opartych na konsumowanych modelach.

Za punkt wyjścia swoich rozważań chcę przyjąć wspomnianą pobieżnie już w poprzednim rozdziale, inspirowaną Peirce'owską semiotyką[1] *triadyczną analizę reprezentacji* (Peirce 1997; por.: Short 2007; Atkin 2010). Z perspektywy Peirce'owskiej triady reprezentowanie zakłada istnienie trzech, odpowiednio ze sobą związanych elementów. Istotę tej idei dobrze oddaje John Haugeland, stwierdzając, że reprezentowanie wymaga: (1) samej reprezentacji, (2) treści reprezentacji oraz (3) roli polegającej na reprezentowaniu czegoś. Jak stwierdza ten autor: „To, co zastępuje coś innego, jest *reprezentacją*; to, co jest zastępowane, stanowi *treść*; a zastępowanie tej treści jest *reprezentowaniem* jej" (Haugeland 1998: 172). Choć uznaję tę charakterystykę za użyteczną, to będę tu stosować nieco innej niż ten autor nomenklatury (zapożyczonej z: von Eckhardt 1993). Proponuję zatem uznać, że reprezentowanie wymaga udziału trzech elementów, do których zaliczają się:

Nośnik reprezentacji – to, co reprezentuje coś innego.
Przedmiot reprezentacji – to, co jest reprezentowane przez nośnik reprezentacji[2].
Interpretacja – wykorzystanie nośnika reprezentacji w funkcji reprezentacji określonego przedmiotu.

O reprezentowaniu możemy mówić jednak dopiero wtedy, gdy te trzy elementy są ze sobą w odpowiedni sposób powiązane. Proponuję przyjąć, że z reprezentowaniem mamy do czynienia, jeśli zostają jednocześnie spełnione dwa warunki:

---

[1] Muszę wyraźnie zaznaczyć, że idee twórcy amerykańskiego pragmatyzmu stanowią dla mnie jedynie inspirację. Przedstawiona tu triadyczna analiza reprezentacji jest jedynie luźno oparta na semiotyce Peirce'a i nie stoją za nią żadne aspiracje o charakterze egzegetycznym.

[2] Termin „przedmiot reprezentacji" jest tu użyty dość ogólnie dla oznaczenia czegokolwiek, co podlega reprezentowaniu. „Przedmiotem" w takim znaczeniu może być nie tylko obiekt, lecz także proces, relacja, stan rzeczy czy zdarzenie.

(W1) Nośnik reprezentacji jest interpretowany jako reprezentacja przedmiotu reprezentacji. Dzięki temu może on pełnić funkcję polegającą na *zastępowaniu* przedmiotu reprezentacji dla kogoś/czegoś.

(W2) Pełnienie przez nośnik reprezentacji roli wymienionej w (W1) jest w systematyczny sposób zależne od relacji zachodzącej między nośnikiem a przedmiotem reprezentacji. To znaczy:

(W2-a) Istnieje pewnego rodzaju relacja, która może zachodzić (lub nie) między nośnikiem a przedmiotem reprezentacji[3].

(W2-b) Zachodzenie relacji wymienionej w (W2-a) to warunek konieczny do tego, by nośnik poprawnie pełnił funkcję wymienioną w (W1). Jeśli relacja taka nie zachodzi, nośnik reprezentacji nie pełni funkcji wymienionej w (W1) w sposób poprawny.

Warunek (W1) dotyczy relacji zachodzącej między nośnikiem oraz interpretacją. Natomiast warunek (W2) dotyczy relacji między nośnikiem a przedmiotem reprezentacji.

Wspomniałem już w poprzednim rozdziale, że inspirowany Peirce'owską semiotyką sposób rozumienia reprezentacji został wykorzystany przez Williama Ramseya (2007: 20–24) w jego analizie naszego przednaukowego czy potocznego pojęcia reprezentacji. Chodzi konkretnie o przednaukowe pojęcie reprezentacji pozamentalnych, czyli takich, które są zewnętrzne względem podmiotu i nie są jako takie jego stanami mentalnymi (por. tabela 2). Naturę tych reprezentacji można, jak się wydaje, dość wygodnie ująć w ramach triady „nośnik–przedmiot–interpretacja". Przyjmowany tu przeze

[3] Zauważmy, że warunek ten nie ogranicza się do sytuacji, w których wymieniona relacja rzeczywiście zachodzi. Podana tu charakterystyka uwzględnia sytuacje, w których odpowiednia relacja między nośnikiem a przedmiotem nie zachodzi, w związku z czym nośnik reprezentacji nie spełnia swojej roli w sposób poprawny (nie zastępuje przedmiotu reprezentacji). W ten sposób dopuszcza się możliwość istnienia reprezentacji błędnych. Kwestia reprezentacji błędnych, jak również ściśle z nią powiązany problem treści reprezentacji, zostaną jeszcze poruszone w dalszej części tego rozdziału (sekcja 4.2.2).

mnie triadyczny sposób rozumienia reprezentacji warto zatem zilustrować, pokazując, jak pozwala on nam „modelować" przednaukowy sposób rozumienia reprezentacji pozamentalnych.

Mówiąc ogólnie, zgodnie z przednaukowym ujęciem reprezentacji zewnętrznych – reprezentacjami mogą być określone zewnętrzne wobec podmiotu (systemu poznawczego) obiekty czy struktury, takie jak słoje w pniu drzewa, mapy albo znaki drogowe. Struktury te możemy uznać za *nośniki* reprezentacji, natomiast to, co przez nie reprezentowane – na przykład wiek drzewa, przestrzenny układ terenu czy fakt występowania ograniczenia prędkości w danym miejscu – za *przedmioty* reprezentacji. Struktury te są reprezentacjami (nośnikami reprezentacji), jeśli są *interpretowane* czy *wykorzystywane jako* reprezentacje. Są one zatem zawsze reprezentacjami *dla* określonych podmiotów; nie ma mowy o reprezentowaniu bez uwikłania w ludzkie praktyki interpretacyjne.

Inspirując się raz jeszcze klasycznymi ideami Charlesa Peirce'a, wśród reprezentacji pozamentalnych możemy wymienić trzy ogólne kategorie: reprezentacje indeksowe, ikoniczne oraz konwencjonalne (por. tabela 2). O przynależności do każdej z nich decyduje rodzaj relacji zachodzącej między nośnikiem a przedmiotem reprezentacji. Jest to relacja, do której odwołuje się warunek (W2-a) na powyższej liście. Reprezentacje indeksowe reprezentują na podstawie związku fizycznego (na przykład relacji przyczynowej) zachodzącego między nośnikiem a przedmiotem reprezentacji. Reprezentacje ikoniczne reprezentują na podstawie podobieństwa zachodzącego między nośnikiem reprezentacji a jej przedmiotem. Reprezentacje konwencjonalne reprezentują na podstawie konwencji, czyli pewnego rodzaju społecznie ustalonej reguły, zgodnie z którą nośnik ma reprezentować określony przedmiot.

Jak już zostało wspomniane, w każdym z tych trzech przypadków reprezentowanie opiera się na tym, że odpowiednie struktury są interpretowane w sposób pozwalający na wykorzystanie ich w roli reprezentacji. Inaczej mówiąc, ludzie mają zdolność do rozumienia, a dzięki temu do używania określonych rzeczy w świecie (nośników reprezentacji) jako reprezentacji określonych obiektów, stanów rzeczy czy zdarzeń (przedmiotów reprezentacji). Reprezentacje

zastępują dla danej osoby to, co reprezentowane, w tym znaczeniu, że pozwalają jej wyciągać określone wnioski czy podejmować decyzje dotyczące przedmiotu reprezentacji bez konieczności wchodzenia z nim w bezpośrednie interakcje[4]. Tego aspektu reprezentacji pozamentalnych dotyczy warunek (W1) na naszej liście.

Należy wreszcie zaznaczyć, że zachodzi ważny związek między obydwiema wymienionymi wyżej własnościami reprezentacji pozamentalnych. Istnieje mianowicie pewna zależność między pełnioną przez nośnik rolą zastępowania czegoś dla kogoś a tym, czy między nośnikiem a przedmiotem reprezentacji zachodzi odpowiednia re-

[4] Niektórzy autorzy (Sloman 2011; Miłkowski 2013: 157–158) krytykują konceptualizowanie roli pełnionej przez reprezentacje w kategoriach zastępowania czegoś dla kogoś (*to stand-in for*). Zwracają oni uwagę na fakt, że jeśli *x* ma zastąpić *y*, to *x* musi być zdolne do odgrywania ról przyczynowych, które są na ogół pełnione przez *y*. Reprezentacje nie pełnią jednak takich samych ról przyczynowych jak to, co jest przez nie reprezentowane. Nie możemy przecież używać reprezentacji w ten sam sposób, w jaki używamy tego, co jest przez nią reprezentowane. Mówiąc obrazowo, nie możemy obrać i zjeść reprezentacji banana (Sloman 2011). Z tych przesłanek autorzy ci wyciągają wniosek, że termin „zastępowanie" nie opisuje poprawnie funkcji pełnionej przez reprezentacje. Wydaje się jednak, iż krytycy interpretują ten termin zbyt dosłownie. Nikt nie twierdzi poważnie, że reprezentacja może dosłownie zastąpić to, co reprezentowane, na przykład tak jak jeden nauczyciel matematyki może zastąpić innego (por. Sloman 2011). Kiedy mowa o reprezentacyjnym „zastępowaniu", chodzi przede wszystkim o to, że reprezentacje zapośredniczają kontakt poznawczy z tym, co reprezentowane. Umożliwiają one podejmowanie decyzji czy wyciąganie wniosków bez konieczności wchodzenia w bezpośrednie interakcje z tym, czego te decyzje czy wnioski dotyczą. Możemy wyciągnąć wniosek o trwającym pożarze, patrząc na unoszący się dym, bez konieczności wchodzenia w kontakt (na przykład percepcyjny) z samym ogniem. Mapa pozwala na wybór optymalnej ścieżki z punktu A do punktu B w sposób zwalniający nas z konieczności wypróbowywania różnych możliwych tras „po omacku". W takim właśnie sensie dym i mapa mogą (poznawczo) „zastąpić" to, co przez nie reprezentowane: mogą być wykorzystywane w określonych celach zamiast samego reprezentowanego obiektu. Co oczywiste, aby zastępować coś w takim sensie, nie trzeba posiadać tych samych własności przyczynowych, co reprezentowany obiekt czy stan rzeczy. (W dalszej części tego rozdziału – sekcja 4.1.4 – powołam się także na inne, nieco mocniejsze rozumienie „zastępowania" czegoś przez reprezentację, odwołujące się do pojęcia rozumowań surogatywnych. Sądzę, że także to rozumienie roli reprezentacji bezproblemowo wpisuje się w metaforę „zastępowania").

lacja (współzmienność/zależność przyczynowa, podobieństwo lub relacja konwencjonalna). Coś może skutecznie pełnić funkcję (nośnika) reprezentacji tylko wtedy, gdy między nim a przedmiotem reprezentacji rzeczywiście zachodzi odpowiedni rodzaj relacji. Dym nie spełni poprawnie swojej roli jako indeksowa reprezentacja ognia, jeśli nie był nim rzeczywiście spowodowany. Mapa albo makieta nie spełni poprawnie roli ikonicznej reprezentacji danego terenu, jeśli go nie odzwierciedla (nie jest do niego podobna pod określonymi względami). Potakujący ruch głową nie będzie poprawnie pełnić roli konwencjonalnej reprezentacji akceptacji (akceptującej postawy), kiedy znajdziemy się we wspólnocie interpretacyjnej, w której gest ten zwyczajowo wyraża przeczenie. Tym samym poprawne spełnienie warunku (W1) wymienionego na powyższej liście staje się możliwe tylko wtedy, gdy rzeczywiście zachodzi relacja, o której mówi warunek (W2-a). Jeśli relacja wymieniona w warunku (W2-a) nie zachodzi, to nośnik nie może poprawnie odgrywać roli polegającej na reprezentowaniu czegoś. Bez istnienia tego rodzaju zależności nie może być w ogóle mowy o reprezentowaniu. To właśnie o tym – czyli o tym, że musi istnieć tego rodzaju zależność między (W1) a (W2-a) – mówi warunek (W2-b).

Jak zatem widzimy, triada „nośnik–przedmiot–interpretacja" stanowi dobre narzędzie pojęciowe służące do konceptualizowania natury przednaukowych reprezentacji pozamentalnych. Pamiętajmy jednak, że właściwym przedmiotem zainteresowania tych rozważań jest możliwość wykorzystania triadycznego rozumienia reprezentacji w kontekście naukowym, czyli w kontekście kognitywistyki i formułowanych w jej ramach wyjaśnień zjawisk poznawczych. Przedmiotem mojego zainteresowania nie są reprezentacje zewnętrzne względem systemu poznawczego, lecz *wewnętrzne*, które mogłyby stanowić *komponenty mechanizmów poznawczych*. Czy Peirce'owska triada okazuje się zatem w jakiś sposób istotna albo przydatna dla projektu poszukiwania koncepcji reprezentacji, które mogłyby wchodzić w skład mechanizmów poznawczych?

Będę bronić tu twierdzenia, że odpowiedź na powyższe pytanie jest twierdząca. Nie wydaje się to stanowiskiem odosobnionym we współczesnej literaturze. Zagadnienie reprezentacji mentalnych

w kontekście Peirce'owskiej triadycznej analizy znaków – nie zawsze jednak *explicite* powołując się na inspiracje Peircem – rozważali chociażby: Barbara von Eckhardt 1993; John Haugeland 1998; Ruth Millikan 2002; Gerard O'Brien, Jon Opie 2004; William Ramsey 2007. Podążam tu więc tropem przetartym już przez różnych autorów. Inspirując się ideami przez nich sformułowanymi, reprezentacje mentalne w sensie istotnym dla kognitywistyki – odgrywające rolę w wyjaśnieniach mechanistycznych zjawisk poznawczych – będę rozpatrywać w kategoriach triady złożonej z nośnika reprezentacji, reprezentowanego przedmiotu i (swoiście rozumianej) interpretacji. Triada ta stanowi teoretyczne rusztowanie dla formułowanej tu koncepcji wyjaśniania reprezentacyjnego w kognitywistyce.

Rzecz jasna opisany do tej pory triadyczny sposób rozumienia reprezentacji powinien zostać odpowiednio „zaadaptowany" do celów kognitywistycznych. Modyfikacje te powinny zatem w szczególności dotyczyć jednego elementu analizy reprezentacji, mianowicie *interpretacji* oraz ściśle z nią związanego warunku (W1). Reprezentacje *poza*mentalne – takie, jak chociażby termometry, mapy, znaki drogowe czy zdania zapisane w języku naturalnym – pełnią funkcję reprezentacji tylko o tyle, o ile są jako takowe interpretowane przez podmioty intencjonalne. Są to zewnętrzne, percepcyjnie dostępne struktury wykorzystywane w roli reprezentacji przez ludzi. Pełnią one funkcję zastępowania przedmiotu reprezentacji tylko dla pełnoprawnych podmiotów intencjonalnych, i tylko o tyle, o ile są włączone w ich praktyki interpretacyjne. Kiedy jednak opisujemy wewnętrzne mechanizmy poznawcze, nie możemy rzecz jasna postulować podmiotów wyposażonych w złożone zdolności interpretacyjne. Myśląc o reprezentacjach w kontekście wewnętrznych mechanizmów odpowiadających za zjawiska poznawcze, musimy przyjąć, że rola zastępowania przez reprezentację tego, co reprezentowane (przedmiotu reprezentacji), będzie realizowana inaczej, niż dzieje się to w przypadku map czy znaków drogowych. Czy oraz jak rola ta może być w takim razie realizowana? Jak i przez kogo/co wewnętrzne reprezentacje mogą być „interpretowane"? Biorąc pod uwagę przyjmowaną w tej książce mechanistyczną perspektywę, znaczenie tych pytań jest zasadnicze. Zajmuje mnie tu przede wszystkim pro-

blem roli, jaką reprezentacje wykonują w mechanizmach poznawczych. Problem dostarczenia takiego „mechanistycznego" rozumienia interpretacji to zaś nic innego, jak problem dostarczenia zgodnej z wymogami mechanicyzmu *funkcjonalnej koncepcji reprezentacji*. Zagadnienie to wyznacza więc sedno prowadzonych tu poszukiwań koncepcji mechanizmów reprezentacyjnych.

Dotychczasowe ustalenia stanowią dobry punkt wyjścia do budowy pozytywnej koncepcji mechanizmów reprezentacyjnych. Zanim będzie można jednak do tego przejść, warto na chwilę nawiązać do wcześniejszych rozważań. Przyjęty tu Peirce'owski sposób postrzegania reprezentacji rzuca bowiem dodatkowe światło na niektóre konkluzje poczynione w poprzednim rozdziale. Podążając za Ramseyem (2007), argumentowałem tam między innymi, że określone sposoby rozumienia reprezentacji w kognitywistyce nie spełniają wymogu opisu zadań, to znaczy przypisują one „reprezentacjom" role funkcjonalne, które *de facto* nie polegają na reprezentowaniu czegoś. Otóż przyjęcie opisanej wyżej, triadycznej perspektywy pozwala zobaczyć dokładniej, na czym polegają braki krytykowanych w poprzednim rozdziale ujęć reprezentacji.

Po pierwsze, przyjrzyjmy się reprezentacjom receptorowym. Istnieje dość oczywisty związek między reprezentacjami receptorowymi a przednaukowo pojmowanymi indeksami. Opiera się on na wspólnym sposobie pojmowania relacji między nośnikiem a przedmiotem reprezentacji. Reprezentacje receptorowe miałyby reprezentować w sposób z grubsza „indeksowy", czyli dzięki współzmienności zachodzącej między nośnikiem reprezentacji a jej przedmiotem. Koncepcja reprezentacji receptorowych dokładnie określa zatem relację między nośnikiem a przedmiotem. Tym samym daje ona odpowiedź na pytanie o to, w jaki sposób reprezentacje tego rodzaju mogłyby spełniać warunek (W2-a). Co więcej, przy wprowadzeniu odpowiednich teleologicznych uzupełnień, można pokazać, że zachodzenie tej relacji (współzmienności) decyduje o tym, czy nośnik poprawnie spełnia swoją funkcję w szerszym systemie. Na przykład kiedy współzmienność między magnetosomem a położeniem północy magnetycznej zostaje w jakiś sposób zaburzona, magnetosom przestaje wykonywać poprawnie swoją „nawigacyjną" rolę w bak-

terii. Reprezentacje receptorowe mogą zatem potencjalnie spełniać warunek (W2-b). Fundamentalny problem z nimi związany dotyczy jednak warunku (W1). Jak widzieliśmy, rola funkcjonalna pełniona w pewnym mechanizmie czy systemie przez (rzekome) reprezentacje oparte na współzmienności nie polega na reprezentacyjnym zastępowaniu czegoś. Tego rodzaju wewnętrzne struktury funkcjonują raczej jako „pośrednicy" przyczynowi między dwiema innymi strukturami, a nie jako reprezentacje. Tym samym warunek (W1) nie jest spełniony w przypadku reprezentacji rozumianych jako receptory.

Po drugie, omawiane w poprzednim rozdziale pojęcie reprezentacji ukrytych wydaje się całkowicie nie przystawać do triadycznej analizy, naruszając zarówno warunek (W1), jak i (W2). Przy takim sposobie rozumienia reprezentacji nie dysponujemy w ogóle ideą jakiegoś osobnego, zlokalizowanego jej nośnika (reprezentacje ukryte mają być wszakże „rozproszone" po całej mechanistycznej strukturze systemu poznawczego). Nie ma zatem jednocześnie mowy o tym, by istniała jakaś relacja między takim wewnętrznym nośnikiem a przedmiotem reprezentacji. Zarówno warunek (W1), jak i (W2) nie może być więc przez (byty postulowane jako) reprezentacje ukryte spełniony.

Po trzecie wreszcie, dotychczasowe ustalenia mogą także wyjaśnić, na czym polega zasadnicza słabość Michaela Andersona i Gregga Rosenberga (2008) koncepcji reprezentacji jako przewodników działań. W kontekście teorii proponowanej przez tych autorów możemy mówić o zlokalizowanych nośnikach reprezentacji. Nośniki te miałyby pełnić w systemie czy mechanizmie rolę polegającą na przewodzeniu działaniami tego systemu czy mechanizmu. Zauważmy jednak, że przedstawiona w poprzednim rozdziale krytyka koncepcji Andersona i Rosenberga opierała się na obserwacji, iż teoria ta jest zbyt szeroka, a przez to podatna na kontrprzykłady. Nie odpowiada ona bowiem na pytanie o to, co odróżnia struktury przewodzące działaniami jako reprezentacje od takich, które realizują tę samą funkcję – to znaczy przewodzą działaniami – nie będąc reprezentacjami (por. Gładziejewski 2015). Mając na uwadze (W1) i (W2), można stwierdzić, że koncepcja Andersona i Rosenberga nie pokazuje, w jaki sposób „przewodnikom działań" udaje się spełnić

oba człony warunku (W2). Niewykluczone, że aby przewodzić działaniami jako reprezentacje, określone wewnętrzne struktury (nośniki) powinny wchodzić w pewnego rodzaju relację z przedmiotem reprezentacji – spełniając w taki sposób warunek (W2-a). Dodatkowo skuteczność takich reprezentacji w pełnieniu roli przewoników działań powinna być systematycznie zależna od tego, czy wchodzą one w tak określoną relację z przedmiotem reprezentacji – co pozwoliłoby im spełnić (W2-b). Anderson i Rosenberg nie poruszają jednak w swojej teorii kwestii relacji między nośnikiem a przedmiotem reprezentacji. Można postawić hipotezę, że właśnie ten fakt czyni ich teorię tak otwartą na kontrprzykłady[5].

4.1.2. Reprezentacje wewnętrzne jako reprezentacje ikoniczne: ustalenia wstępne

Wróćmy teraz do głównego nurtu rozważań. Pragnę tu wykorzystać triadę „nośnik–przedmiot–interpretacja" jako rusztowanie, na którym można zbudować zgodne z mechanicyzmem, eksplanacyjnie wartościowe pojęcie reprezentacji mentalnych. Wspomniałem już, że triada ta dobrze sprawdza się przy analizie potocznego pojęcia reprezentacji pozamentalnych. Jak jednak zauważa Ramsey (2007: 22), *mentalne* reprezentacje, na które powołują się kognitywiści i filozofowie, okazują się czasem „mechanicznymi", zinternalizowanymi wersjami reprezentacji *pozamentalnych* (zewnętrznych). Na przykład, jak już zaznaczyłem, kognitywistyczne pojęcie reprezentacji receptorowych wydaje się oparte na (przednaukowym) pojęciu reprezentacji indeksowych. Być może eksplanacyjnie wartościowe reprezentacje, z których mogą korzystać wewnętrzne mechanizmy poznania, to rzeczywiście reprezentacje nietrywialnie przypomina-

[5] Uprzedzając nieco przebieg dalszych rozważań, warto powiedzieć, że zgodnie z bronioną tu koncepcją reprezentacje to przewodniki działań, które są zarazem wewnętrznymi modelami tego, co reprezentowane. Podkreślenie roli eksplanacyjnej podobieństwa zachodzącego między nośnikiem a przedmiotem reprezentacji (między modelem a tym, co przezeń modelowane) jest według mnie właśnie kluczowym brakującym elementem w teorii Andersona i Rosenberga (por. Gładziejewski 2015).

jące funkcjonalnie swoje „zdroworozsądkowe", przednaukowe odpowiedniki pozamentalne. Spróbuję potraktować tę ideę zupełnie poważnie, jako heurystyczny „trop" w poszukiwaniach koncepcji mechanizmów reprezentacyjnych.

Wspomniałem powyżej, że reprezentacje pozamentalne można podzielić na trzy kategorie: indeksowe, ikoniczne i konwencjonalne. Jak już wiemy, dla bieżących celów teoretycznych nie nadają się te pierwsze. „Wywodzące" się z nich reprezentacje receptorowe nie zasługują na status reprezentacji pod względem pełnionych ról funkcjonalnych – nie spełniają wymogu opisu zadań. Reprezentacje indeksowe nie stanowią więc dobrego źródła inspiracji dla koncepcji mechanizmów reprezentacyjnych. Warto jednak zapytać, co z pozostałymi dwiema kategoriami reprezentacji pozamentalnych? Czy one „rokują" lepiej jako potencjalna podstawa, na której można owocnie wypracować naukowe pojęcie reprezentacji, a w związku z tym także koncepcję mechanizmów reprezentacyjnych?

Spójrzmy najpierw na reprezentacje konwencjonalne. Jak już zostało zaznaczone, w ich przypadku relacja, której dotyczy warunek (W2-a) – czyli ta zachodząca między nośnikiem a przedmiotem reprezentacji – ma naturę konwencjonalną. Nośnik nie musi być ani podobny do przedmiotu reprezentacji, ani nie musi wchodzić z nim w żaden związek przyczynowy. Zamiast tego powinna istnieć pewnego rodzaju konwencjonalna reguła określająca, co stanowi przedmiot reprezentowany przez określony nośnik. W ten właśnie sposób znak drogowy może reprezentować zasady pierwszeństwa na skrzyżowaniu, a potakujący ruch głową – akceptację czegoś. Bardzo szybko okazuje się jednak, że w kognitywistyce takiego rodzaju relacja między nośnikiem a przedmiotem nie może wchodzić w grę. Mówiąc z grubsza, konwencje związane ze znakami drogowymi i potakującymi ruchami głową są ustalane społecznie, to znaczy przez zbiorowości podmiotów intencjonalnych. Istnienia tych ostatnich nie możemy postulować, kiedy przedmiotem naszego zainteresowania jest wewnętrzna, mechanistyczna organizacja systemu poznawczego (von Eckhardt 1993: 206, 234–239). Skoro tak, nie ma sensu mówienie o istnieniu wewnątrz systemu poznawczego bytów, które

mogłyby ustalać konwencjonalne reguły[6]. Jeśli poszukujemy pojęcia reprezentacji pełniącego rolę w mechanistycznych wyjaśnieniach formułowanych przez kognitywistów, nie możemy go oprzeć na pojęciu reprezentacji konwencjonalnych.

Z trzech wcześniej wyróżnionych kategorii reprezentacji pozamentalnych, na podstawie których potencjalnie moglibyśmy oprzeć koncepcję mechanizmów reprezentacyjnych, pozostała tylko jedna: reprezentacje *ikoniczne*. Wydaje się, że w nich wreszcie można znaleźć oparcie dla rozwijanego tu projektu. Chcę bronić idei, że reprezentacje spełniające rzeczywiście taką rolę w mechanizmach poznawczych pod istotnymi względami przypominają funkcjonalnie właśnie (przednaukowe, pozamentalne) reprezentacje ikoniczne. Pozostała część tego rozdziału będzie poświęcona rozwinięciu i obronie tezy, że naukowe, eksplanacyjnie wartościowe rozumienie reprezentacji może zostać oparte właśnie na przednaukowym pojęciu reprezentacji ikonicznych. Zacznijmy od precyzyjniejszego scharakteryzowania tych ostatnich.

---

[6] Gerard O'Brien i Jon Opie (2004) zwracają uwagę, że istnieje pewna potencjalna droga wyjścia z tego problemu. Otóż możemy uznać, że w kontekście kognitywistyki pojęcie „konwencji" przyjmuje inne znaczenie – takie, które nie naraża nas na popełnienie błędu homunkularnego. W naukach kognitywnych stosowne konwencje mogą być rozumiane jako reguły (instrukcje) zachodzących w systemie procesów obliczeniowych (przy czym reguły te nie muszą być jako takie wewnętrznie reprezentowane w systemie). Ci sami autorzy twierdzą jednak, że taki zabieg nie może się powieść. Dlaczego? Otóż zakładają oni, że konwencjonalna relacja między nośnikiem a przedmiotem reprezentacji powinna w założeniu *wpływać na* czy też *determinować* proces „interpretacji" nośnika. Jest to w ich koncepcji odpowiednik, w przybliżeniu, postulowanego tu przeze mnie wcześniej warunku (W2-b). O'Brien i Opie (2004) pokazują, że taka „konwencjonalna" determinacja procesu interpretacji jest niemożliwa, jeśli przez „konwencje" rozumiemy reguły czy instrukcje obliczeniowe. Konwencje rozumiane jako reguły czy instrukcje obliczeniowego przetwarzania informacji jako takie nie wpływają na sposób interpretacji nośnika, lecz *konstytuują* one tę interpretację (sposób, w jaki nośnik jest używany w ramach systemu poznawczego). Być „interpretowanym" znaczy w takiej perspektywie to samo, co bycie przetwarzanym w określony sposób. O istnieniu jakiejś osobnej relacji między nośnikiem a przedmiotem reprezentacji, determinującej sposób funkcjonowania nośnika w mechanizmie, nie może być w takiej perspektywie mowy (O'Brien, Opie 2004).

Reprezentacje ikoniczne opierają się na podobieństwie. Mówiąc bardzo ogólnie, oznacza to, że podzielają one niektóre własności z tym, co jest przez nie reprezentowane – i w ten sposób odzwierciedlają to ostatnie. W reprezentacjach tego rodzaju to właśnie podobieństwo stanowi relację między nośnikiem a przedmiotem reprezentacji, o której mówi wymieniony wcześniej warunek (W2-a). Oto kilka przykładów reprezentacji ikonicznych:

Mapa przestrzenna – przestrzenny układ mapy (na przykład układ przedstawionych na niej punktów i linii) odzwierciedla układ przestrzenny określonego terenu (na przykład budynków i dróg w obrębie pewnego miasta).
Zdjęcie – dwuwymiarowy obraz, który odzwierciedla pewne aspekty przedstawianej sceny (na przykład kolory, kształt i rozmieszczenie obiektów czy ekspresje emocjonalne osób przedstawionych).
Ruchoma makieta – relatywna wielkość, wzajemne odległości i relatywny sposób poruszania się obiektów na makiecie odzwierciedla relatywną wielkość, wzajemne odległości i relatywny sposób poruszania się reprezentowanych obiektów (na przykład planet Układu Słonecznego).
Próbka koloru – rodzaj i odcień koloru próbki jest ten sam, co rodzaj i odcień koloru reprezentowanego (na przykład koloru farby, którą chcemy wykorzystać do pomalowania ścian pokoju).

Rzecz jasna nie jest tak, że samo zachodzenie podobieństwa gwarantuje czemuś status reprezentacji. Wszystkie powyższe przykłady reprezentacji ikonicznych są nimi także ze względu na realizowane przez siebie *funkcje*. Inaczej mówiąc, nie ma reprezentacji ikonicznych bez podmiotów, które rozpoznają je jako reprezentacje i jako takie je też wykorzystują. Mapy, makiety, zdjęcia i próbki kolorów to kulturowe artefakty uwikłane w ludzkie praktyki interpretacyjne, i tylko dzięki istnieniu tych praktyk możemy powiedzieć, że reprezentują one cokolwiek. W każdym z wyżej wymienionych przypadków funkcja reprezentowania opiera się na tym, że istnieją istoty, które poznawczo wykorzystują relację podobieństwa zachodzącą

między nośnikiem a przedmiotem reprezentacji: między strukturą mapy a strukturą terenu, między kolorem próbki a kolorem farby i tak dalej. Ten akt wykorzystania podobieństwa stanowi element reprezentacyjnej triady, nazwany tu wcześniej „interpretacją".

Powyższe uwagi pozwalają dookreślić cel tego rozdziału. Poszukuję tu koncepcji reprezentacyjnego wyjaśnienia mechanistycznego, czyli wyjaśnienia odwołującego się do mechanizmu reprezentacyjnego. O reprezentacyjnym statusie mechanizmu ma z kolei decydować – zgodnie z ustaleniami poczynionymi w poprzednim rozdziale – fakt, że mechanizm tego rodzaju posiada komponent, którego funkcja (wykonywana operacja) polega na reprezentowaniu czegoś. Otóż chcę tu rozwinąć i obronić ideę, że funkcja komponentu zajmującego się w ramach mechanizmu reprezentowaniem czegoś przypomina w jakimś nietrywialnym sensie funkcję pełnioną przez pozamentalne reprezentacje ikoniczne. Mówiąc bardziej obrazowo, pozostała część tego rozdziału zostanie poświęcona rozwinięciu oraz obronie tezy, zgodnie z którą mechanistyczne wyjaśnienie pewnego zjawiska poznawczego za pomocą reprezentacji to wyjaśnienie go za pomocą mechanizmu posiadającego komponent pełniący funkcję swoistej wewnętrznej, mechanicznej „mapy" czy „makiety".

Projekt oparcia naukowego, eksplanacyjnie wartościowego pojęcia reprezentacji na pojęciu reprezentacji ikonicznych natrafia jednak *prima facie* na dwa fundamentalne problemy. Po pierwsze, jest całkowicie nieoczywiste, czy (ewentualnie, jaki) sens możemy nadać idei, jakoby wewnętrzne reprezentacje – reprezentacje rozumiane jako komponenty wewnętrznych mechanizmów poznawczych – były *podobne* do tego, co reprezentowane (por.: Cummins 1989: 31––32; O'Brien, Opie 2004). Jak zauważają Gerard O'Brien i Jon Opie, „nie ma nic bardziej oczywistego niż fakt, że nasze umysły zdolne są do reprezentowania aspektów świata, które nie mogą zostać odtworzone [*replicated*] w tkance nerwowej" (2004: 9). Ujmując tę myśl jeszcze bardziej dosadnie: jak stworzyć koncepcję wewnętrznych reprezentacji opartych na podobieństwie między nośnikiem a przedmiotem reprezentacji, która nie implikowałaby szeregu absurdalnych konsekwencji – na przykład, że reprezentowanie banana wymaga posiadania „w głowie" obiektów mających fizyczne wła-

sności (choćby kolor czy kształt) bananów? Nazwijmy ten problem skrótowo „problemem relacji podobieństwa".

Po drugie, nawet jeśli założymy, że problem relacji podobieństwa może być rozwiązany, nadal trzeba odpowiedzieć na pytanie o to, czy oraz w jaki sposób nasze „mechaniczne" reprezentacje ikoniczne realizują *role funkcjonalne* w jakimś sensie zbliżone czy analogiczne do ról, które są odgrywane przez mapy, makiety, zdjęcia czy próbki kolorów. Co by to znaczyło dla takiej *wewnętrznej* reprezentacji, że pełni ona rolę polegającą na *zastępowaniu* czegoś? Czy jest to w ogóle możliwe bez postulowania osób (podmiotów) interpretujących takie reprezentacje? Jak „zmechanizować" proces interpretacji? Pytania te dotyczą ogólnego zagadnienia, które możemy nazwać „problemem roli funkcjonalnej"[7]. Rozwiązanie go wymaga takiego scharakteryzowania roli pełnionej przez wewnętrzne, mechanistycznie rozumiane reprezentacje, by jednocześnie (1) przypisanie tej roli wewnętrznej reprezentacji nie wiązało się z popełnieniem błędu homunkularnego; (2) wewnętrzne struktury odgrywające tę rolę czyniły zadość Ramseyowskiemu wymogowi opisu zadań (to znaczy rzeczywiście pełniły w mechanizmie funkcję polegającą na reprezentowaniu czegoś). Jak zobaczyliśmy w poprzednim rozdziale, właśnie na tym etapie zawodzą próby oparcia kognitywistycznego pojęcia reprezentacji na przednaukowym pojęciu reprezentacji indeksowych. Wewnętrzne struktury, których funkcjonowanie opiera się na współzmienności, nie funkcjonują tak jak zewnętrzne reprezentacje indeksowe (por. Ramsey 2007: 118–150). Czy koncepcja postulująca istnienie wewnętrznych odpowiedników zewnętrznych reprezentacji ikonicznych daje lepsze rezultaty?

---

[7] Warto zauważyć, że problem podobieństwa oraz problem roli funkcjonalnej nie są w istocie do końca odrębne i niezależne. Trzeba wszakże pamiętać o wspomnianym wyżej warunku (W2-b): od relacji zachodzącej (lub nie) między nośnikiem a przedmiotem reprezentacji powinno zależeć to, czy nośnik poprawnie wykonuje swoją funkcję. Potrzeba zatem takiej funkcjonalnej koncepcji reprezentacji, która jest spójna z ideą, iż powinien zachodzić określony związek między, z jednej strony, rolą pełnioną przez nośnik reprezentacji (komponent stanowiący taki nośnik) w mechanizmie a, z drugiej strony, podobieństwem zachodzącym na linii nośnik–przedmiot reprezentacji.

Celem następnych dwóch sekcji tego podrozdziału będzie rozwiązanie problemu relacji podobieństwa (sekcja 4.1.3) oraz problemu roli funkcjonalnej (sekcja 4.1.4). Będzie to jednoznaczne z pokazaniem, że inspirując się zewnętrznymi ikonami, można wypracować pojęcie reprezentacji wewnętrznych, które jest zgodne z wymogami mechanicyzmu i eksplanacyjnie wartościowe dla kognitywistyki. Zaproponowane tu rozwiązania będą też stanowić podstawę dla prezentowanej w podrozdziale 4.2 koncepcji mechanizmów reprezentacyjnych jako takich, których działanie opiera się na konsumowanych modelach.

4.1.3. Wewnętrzne reprezentacje ikoniczne: problem relacji podobieństwa. Pojęcie reprezentacji strukturalnych

Zacznijmy od problemu relacji podobieństwa. Nie jest on tak poważny, jak może się początkowo wydawać. Idei, że wewnętrzne nośniki reprezentacji – zlokalizowane, przynajmniej kiedy mówimy o ludzkim systemie poznawczym, w ośrodkowym układzie nerwowym – są podobne do tego, co reprezentują, można nadać wiarygodny sens. Za Gerardem O'Brienem i Jonem Opie (2004) należałoby bowiem odróżnić dwa rodzaje relacji podobieństwa. Pierwszy z nich to *podobieństwo pierwszego rzędu*. Opiera się ono na tym, że dwa obiekty – w naszym wypadku: nośnik i przedmiot reprezentacji – podzielają ze sobą jakieś własności fizyczne, takie jak kolor, kształt czy wielkość. W ten właśnie sposób próbka koloru może być podobna do (reprezentowanej) farby: podzielają one własność polegającą na posiadaniu określonego koloru. Wydaje się, że pozorna absurdalność idei, jakoby nośniki wewnętrznych, mentalnych reprezentacji mogły reprezentować na podstawie podobieństwa, bierze się z założenia, iż chodzi o podobieństwo pierwszego rzędu. To właśnie przyjęcie takiego rozumienia podobieństwa kazałoby nam sądzić, że wewnętrzny nośnik dosłownie podziela z przedmiotem reprezentacji określone własności fizyczne.

Sytuacja zmienia się jednak, jeśli odróżnimy podobieństwo w opisanym wyżej znaczeniu od relacji, którą O'Brien i Opie (2004) nazywają *podobieństwem drugiego rzędu* (por. także: Palmer 1979;

Cummins 1989: 85–102; Swoyer 1991; Blachowicz 1997; Bartels 2006; Braddon-Mitchell, Jackson 2007: 188–193; Ramsey 2007: 77––79). Podobieństwo drugiego rzędu to inaczej podobieństwo *strukturalne*. Opiera się ono na analogii czy korespondencji zachodzącej na poziomie struktury dwóch złożonych obiektów. W kontekście problemu reprezentacji chodzi tu rzecz jasna o analogię czy korespondencję zachodzącą między strukturami nośnika i przedmiotu reprezentacji. Jeśli przyjmiemy takie rozumienie relacji podobieństwa, powinniśmy zarówno o nośniku, jak i przedmiocie reprezentacji myśleć jako o bytach ustrukturyzowanych, to znaczy złożonych z elementów powiązanych ze sobą określonymi relacjami. Podobieństwo strukturalne zachodzące między nośnikiem a przedmiotem polegałoby na tym, że układ relacyjny nośnika odzwierciedla układ relacyjny przedmiotu reprezentacji. Jak to ujmuje Christopher Swoyer, w reprezentacjach opartych na podobieństwie strukturalnym „układ relacji zachodzących między częściami składowymi reprezentowanego zjawiska jest odzwierciedlany [*mirrored*] przez układ relacji zachodzących między częściami składowymi samej reprezentacji" (1991: 452).

Nazwijmy zbiorczo reprezentacje oparte na podobieństwie strukturalnym „reprezentacjami strukturalnymi", a dla uproszczenia: „S-reprezentacjami"[8]. Kategoria ta ma obejmować zarówno niektóre zewnętrzne, pozamentalne reprezentacje ikoniczne[9], jak i reprezentacje wewnętrzne, mentalne – o ile takie w ogóle istnie-

[8] Termin „S-reprezentacja" został pierwotnie zastosowany przez Cumminsa (1989: 96–97) dla oznaczenia reprezentacji opartych na procesie symulacji. Przyjmuję tu (por. dalsza część tej sekcji), że symulacja to proces opierający się na zachodzeniu dynamicznego (diachronicznego) podobieństwa strukturalnego między nośnikiem a przedmiotem reprezentacji. W tej pracy terminem „S-reprezentacja" będę się posługiwać dla oznaczenia zarówno reprezentacji opartych na statycznym (synchronicznym), jak i dynamicznym (diachronicznym) podobieństwie strukturalnym. Innymi słowy, tym, co istotne dla S-reprezentacji w przyjętym tu znaczeniu, jest (jedynie) fakt, iż reprezentacje tego rodzaju są oparte na podobieństwie strukturalnym.

[9] Chodzi tu o te reprezentacje ikoniczne, w których przypadku relacja zachodząca między nośnikiem a przedmiotem reprezentacji to podobieństwo drugiego, a nie pierwszego rzędu.

ją. Wśród tak ogólnie rozumianych S-reprezentacji warto wyróżnić dwie subkategorie (O'Brien, Opie 2004). Z jednej strony mamy S-reprezentacje, w których podobieństwo występujące między nośnikiem a przedmiotem reprezentacji zachodzi na poziomie struktur tego samego rodzaju. Na przykład w mapach przestrzennych struktura przestrzenna (metryczna czy topograficzna) mapy odzwierciedla strukturę przestrzenną reprezentowanego terenu. Relatywne wzajemne położenia i odległości między elementami mapy odpowiadają relatywnym wzajemnym położeniom i odległościom między elementami terenu. Załóżmy więc, że pewnym budynkom *A*, *B* i *C* odpowiadają na mapie, odpowiednio, punkty *A'*, *B'* i *C'*. Zakładając, że mapa adekwatnie odzwierciedla teren, możemy na przykład orzec, że jeśli punkt *A'* znajduje się bliżej punktu *B'* niż punktu *C'*, to budynek *A* znajduje się bliżej budynku *B* niż *C*; jeśli *A'* znajduje się między *B'* a *C'*, to budynek *A* znajduje się między budynkami *B* a *C* i tak dalej.

Z drugiej strony możemy wyróżnić S-reprezentacje, w których struktura nośnika jest *innego rodzaju* niż reprezentowana przez nią struktura przedmiotu reprezentacji (O'Brien, Opie 2004). W takim przypadku nośnik i przedmiot nie muszą podzielać ze sobą żadnych własności fizycznych: zachodzi między nimi podobieństwo drugiego rzędu (strukturalne), a jednocześnie w ogóle nie zachodzi podobieństwo pierwszego rzędu (fizyczne). Na przykład przestrzenna struktura strzałek i napisów na diagramie stanowiącym drzewo genealogiczne danej rodziny może odzwierciedlać zachodzącą w tej ostatniej strukturę pokrewieństwa. Jeśli dwa elementy takiego diagramu są ze sobą połączone strzałką danego typu, to między określonymi (odpowiadającymi tym elementom diagramu) członkami rodziny zachodzi określona (odpowiadająca tego typu strzałce) relacja pokrewieństwa. Widząc, że dwa imiona na diagramie są połączone podwójną strzałką, możemy w ten sposób dla przykładu orzec, że osoby noszące te imiona znajdują się w relacji „bycia rodzeństwem". Diagram nie przypomina reprezentowanej rodziny fizycznie. Podobieństwo między nimi może występować tylko na poziomie struktur czy układów relacyjnych.

Choć powyższe uwagi pozwalają zrozumieć naturę podobieństwa strukturalnego na poziomie intuicyjnym, warto spróbować wyrazić ją nieco bardziej precyzyjnie. Jak już wspomniałem, nośnik S-reprezentacji należy rozumieć jako system czy układ elementów wchodzących ze sobą w określonego rodzaju relacje. Przyjmijmy więc, że nośnik reprezentacji to system *Sn* = (*N*, *Rn*), gdzie *N* to zbiór elementów składowych nośnika, a *Rn* to zbiór relacji mogących zachodzić między elementami zbioru *N* (tu i dalej za: O'Brien, Opie 2004; por. też: Swoyer 1991; Bartels 2006). Przyjmijmy też, że przedmiot reprezentacji to również złożony system czy układ, *Sp* = (*P*, *Rp*), gdzie *P* to zbiór elementów składowych przedmiotu, a *Rp* to zbiór relacji, jakie mogą zachodzić między elementami zbioru *P*. Otóż *Sn* = (*N*, *Rn*) i *Sp* = (*P*, *Rp*) są do siebie strukturalnie podobne, jeśli dla co najmniej niektórych elementów z *N* oraz co najmniej niektórych relacji z *Rn*: (1) istnieje funkcja wzajemnie jednoznaczna przypisująca elementy *N* elementom *P* oraz funkcja wzajemnie jednoznaczna przypisująca relacje z *Rn* relacjom z *Rp*; (2) jest tak, że jeśli jakaś relacja z *Rn* zachodzi między poszczególnymi elementami *N*, to odpowiadająca relacja z *Rp* zachodzi między odpowiadającymi elementami z *P*. Jeśli spełniony zostaje warunek (2), możemy powiedzieć, że elementy składowe *Sn* i zachodzące między nimi relacje mogą być przypisane elementom i relacjom w *Sp* w sposób *odzwierciedlający strukturę Sp*.

Niektórzy autorzy odwołujący się do S-reprezentacji stawiają mocny wymóg, by między nośnikiem a przedmiotem reprezentacji zachodziło kompletne podobieństwo strukturalne, to znaczy, by nośnik i przedmiot reprezentacji były ze sobą *izomorficzne* (por.: Palmer 1979; Cummins 1989: 103; Ramsey 2007: 77–78). Zgodnie z tym wymogiem powinno być tak, że każdemu elementowi i każdej relacji w nośniku jest przypisany (w sposób odzwierciedlający strukturę) dokładnie jeden element (jedna relacja) w przedmiocie reprezentacji, a także *vice versa*: każdemu elementowi i każdej relacji w przedmiocie reprezentacji jest przypisany (w sposób odzwierciedlający strukturę) dokładnie jeden element (jedna relacja) w nośniku. Inni autorzy uznają taki warunek za zbyt mocny i nierealistyczny, ponieważ w przypadku większości S-reprezentacji przedmiot repre-

zentacji okazuje się „bogatszy" strukturalnie niż jej nośnik (Swoyer 1991; Bartels 2006; Miłkowski 2013: 149–151). Na przykład mapa stanowi dwuwymiarową reprezentację trójwymiarowego terenu. W związku z tym autorzy ci skłaniają się ku charakteryzowaniu relacji między nośnikiem a przedmiotem reprezentacji w kategoriach homomorfizmu (Bartels 2006) bądź tak zwanego izomorfizmu osadzonego (*embedded isomorphism*; Swoyer 1991). W tym pierwszym przypadku (przy powoływaniu się na homomorfizm) przyjmuje się, że funkcja przypisująca elementy/relacje w nośniku elementom/relacjom w przedmiocie reprezentacji nie jest wzajemnie jednoznaczna[10]. W przypadku drugim (przy powoływaniu się na izmorfizm osadzony) dopuszcza się z kolei możliwość, że przedmiot reprezentacji zawiera elementy i relacje, które w ogóle nie mają swoich odpowiedników w nośniku reprezentacji.

Nawet uznanie homomorfizmu lub izomorfizmu osadzonego za podstawę S-reprezentacji wiąże się jednak z pewnymi problemami. W przypadku zewnętrznych, pozamentalnych S-reprezentacji często mamy wszakże do czynienia z sytuacją, w której reprezentacja spełnia poprawnie swoją funkcję pomimo faktu, że struktura nośnika nie odzwierciedla struktury reprezentowanego przedmiotu do końca wiernie czy adekwatnie. Nośnik może zatem zawierać elementy lub relacje, które w ogóle nie mogą być przypisane w sposób odzwierciedlający strukturę elementom (relacjom) w przedmiocie. Jak się na przykład wydaje, mapa może poprawnie reprezentować określony teren nawet wtedy, gdy jej przestrzenna struktura stanowi jedynie „przekłamane", uproszczone czy wyidealizowane odzwierciedlenie struktury tego terenu. Trzeba jednak podkreślić, że pojęcie „podobieństwa strukturalnego" ma być na tyle szerokie, by obejmować sobą nie tylko izomorfizm (w tym izomorfizm osadzony) i homomorfizm, ale też właśnie tego rodzaju przypadki (O'Brien, Opie 2004; por. Giere 2004). Wszakże zgodnie z podaną wyżej definicją odzwierciedlająca strukturę odpowiedniość musi zachodzić przy-

[10] Oznacza to, że jednemu elementowi/relacji nośnika może odpowiadać kilka (a nie dokładnie jeden) elementów/relacji w przedmiocie.

najmniej dla *niektórych elementów* nośnika reprezentacji i przynajmniej *niektórych relacji* zachodzących między tymi elementami.

Przyjęcie tak osłabionego rozumienia relacji podobieństwa między nośnikiem a przedmiotem reprezentacji w naturalny sposób rodzi jednak pytanie: jak bardzo podobny powinien być nośnik reprezentacji do jej przedmiotu, by można było powiedzieć, że pierwszy poprawnie reprezentuje drugi[11]? Otóż jeśli analizujemy S-reprezentacje przez pryzmat Peirce'owskiej triady, odpowiedź wydaje się jasna: w zasadniczym stopniu zależy to od *interpretacji*. Żeby stanowić poprawną reprezentację, nośnik musi być na tyle podobny do przedmiotu reprezentacji, aby mógł on być w udany czy poprawny sposób interpretowany, a dzięki temu *używany* jako reprezentacja. Mówiąc inaczej, nośnik musi być do przedmiotu podobny w takim stopniu, jaki jest konieczny, aby mógł on skutecznie odgrywać dla kogoś rolę polegającą na zastępowaniu przedmiotu reprezentacji. Nawet niekompletna, uproszczona mapa może poprawnie reprezentować teren wtedy, gdy wystarcza ona komuś, aby podejmować poprawne (pożądane ze względu na realizowane cele) decyzje dotyczące tego, jak poruszać się w obrębie danego terenu[12]. Na przykład mapa (schemat) metra może stanowić dobrą podstawę do planowania podróży nawet w sytuacji, gdy nie odzwierciedla on przebiegu linii czy relatywnego położenia stacji w sposób całkiem precyzyjny. Trzeba zatem pamiętać, że reprezentacje są zawsze „reprezentacjami dla". Po-

[11] Pytanie to należy odróżnić od bardziej ogólnego pytania o to, kiedy jakiś nośnik w ogóle reprezentuje jakiś przedmiot, niekoniecznie poprawnie. W przypadku S-reprezentacji pozamentalnych odpowiedź na to pytanie brzmi: wtedy, gdy jest on przez kogoś interpretowany, a przez to używany jako reprezentacja jakiegoś przedmiotu. W przypadku S-reprezentacji wewnętrznych odpowiedź brzmi: nośnik reprezentuje coś, o ile jest on wykorzystywany jako reprezentacja w ramach szerszego mechanizmu czy systemu. Na czym jednak to „wykorzystywanie" polega? Odpowiedź zostanie sformułowana w toku dalszych rozważań prowadzonych w tym rozdziale (w sekcji 4.1.4).

[12] Zauważmy na marginesie, że prowadzi to do wniosku, iż nie sposób mówić o poprawności lub niepoprawności S-reprezentacji *simpliciter*. S-reprezentacje można w ten sposób kwalifikować tylko relatywnie do ich użytkowników wraz z ich praktycznymi interesami. Mapa miasta w zupełności poprawna ze względu na potrzeby nawigacyjne turysty może okazać się beznadziejnie nieprecyzyjna, a przez to bezużyteczna dla architekta planującego budowę osiedla.

jęcie interpretacji, ściśle powiązane z ideą użytkownika reprezentacji (tego, dla kogo stanowi ona reprezentację), jest tu nieodzowne.

Pamiętajmy również jednak, że mapa to reprezentacja publiczna, pozamentalna, używana przez pełnoprawne podmioty intencjonalne. Czy analogiczny sposób myślenia można zastosować w kontekście S-reprezentacji mentalnych, takich, które mogłyby pełnić funkcje w wewnętrznych mechanizmach poznawczych? Czy w przypadku takich reprezentacji o zakresie strukturalnego podobieństwa potrzebnego do bycia poprawną reprezentacją także decydują kwestie związane z „interpretacją" czy też użyciem? Sądzę, że odpowiedź jest twierdząca. Konceptualne zasoby potrzebne, aby uzasadnić taką odpowiedź, postaram się wypracować w dalszej części tego rozdziału (w podrozdziale 4.2). Póki co można jednak zostawić ten problem na boku.

Dotychczasowe rozważania nie brały pod uwagę temporalnego czy procesualnego aspektu S-reprezentacji. Tymczasem warto podkreślić, że S-reprezentacje mogą być reprezentacjami o charakterze *dynamicznym*. Nie muszą się one zawężać do reprezentowania pojedynczych „klatek" czasowych. Mogą mieć także naturę dynamiczną (diachroniczną), a dzięki temu reprezentować zmiany czy procesy, jakim podlega (lub może podlegać) przedmiot reprezentacji (por. Swoyer 1991). Na przykład w ruchomej makiecie Układu Słonecznego sposób, w jaki jej elementy zmieniają w czasie swoje relatywne położenia, odzwierciedla sposób, w jaki zmieniają się w czasie relatywne położenia Słońca i poruszających się wokół niego planet. W tym oraz podobnych przypadkach możemy potraktować nośnik reprezentacji jako strukturę podlegającą zmianom w kolejnych momentach czasowych $Tn = (tn_1, tn_2, ..., tn_n)$. Tak samo możemy ująć przedmiot reprezentacji: to struktura zmieniająca się w momentach czasowych $Tp = (tp_1, tp_2, ..., tp_n)$. S-reprezentacja jest dynamiczna (diachroniczna) wtedy, gdy (1) kolejnym momentom czasowym $(tn_1, tn_2, ..., tn_n)$ nośnika możemy przyporządkować kolejne momenty czasowe $(tp_1, tp_2, ..., tp_n)$ przedmiotu reprezentacji, w taki sposób, że (2) w dowolnym momencie $(tn_1, tn_2, ..., tn_n)$ nośnik będzie strukturalnie podobny do przedmiotu reprezentacji w odpowiadającym (przyporządkowanym mu) momencie czasowym (tp_1,

tp2,..., tpn). Można więc powiedzieć, że dynamiczna S-reprezentacja to czasowo uporządkowana sekwencja synchronicznych podobieństw strukturalnych między nośnikiem a przedmiotem reprezentacji. Możemy utożsamić dynamiczną S-reprezentację z procesem *symulacji*[13].

Zanim będę mógł przejść dalej, muszę jasno określić rolę, jaką pełni pojęcie podobieństwa strukturalnego przy rozwiązywaniu problemu relacji podobieństwa. Jak pamiętamy, zagadnienie to da się streścić w pytaniu: czy można utrzymywać, że wewnętrzne reprezentacje są oparte na podobieństwie (między nośnikiem a przedmiotem reprezentacji), unikając zarazem trywialnie fałszywych, wręcz absurdalnych konsekwencji takiego stanowiska (na przykład przekonania, że reprezentowanie banana wymaga posiadania „w głowie" żółtego, podłużnego przedmiotu)? Odpowiedź na to pytanie brzmi: tak, można utrzymywać takie stanowisko, lecz pod warunkiem że chodzi o podobieństwo drugiego rzędu (strukturalne). Jeśli mówimy o podobieństwie strukturalnym, nośnik i przedmiot nie muszą dzielić ze sobą fizycznych własności. Dowolny rodzaj struktury może reprezentować dowolny inny rodzaj struktury, o ile między obydwoma zachodzi podobieństwo na poziomie układu relacji; podobieństwo pierwszego rzędu (fizyczne) nie musi w ogóle towarzyszyć podobieństwu drugiego rzędu (O'Brien, Opie 2004). Oparcie pojęcia reprezentacji mentalnych na relacji podobieństwa strukturalnego nie pociąga absurdalnych, trywialnie fałszywych czy choćby wysoce nieprawdopodobnych konsekwencji. Nie oznacza to rzecz jasna, że unikamy w ten sposób innych trudnych pytań. Jakie rodzaje struktur w systemie poznawczym stanowią nośniki reprezentacji? Czy chodzi o struktury, które możemy odnaleźć na poziomie po-

[13] Warto zwrócić uwagę, że jest to szerokie rozumienie symulacji, które nie zakłada, iż symulacje są z konieczności procesami obliczeniowymi. Symulacje obliczeniowe są bez wątpienia ciekawą subkategorią symulacji w ogóle. Jednak zgodnie z bronionym tu ujęciem każdy rodzaj diachronicznego podobieństwa strukturalnego może stanowić – o ile jest wykorzystywany w funkcji reprezentacji – symulację. Proces, jakiemu podlega ruchoma makieta Układu Słonecznego, będzie więc symulacją, tak samo jak przeprowadzana na komputerze obliczeniowa symulacja procesów pogodowych.

tencjałów wywołanych w pojedynczych neuronach, czy może raczej chodzi o struktury czy wzory aktywności całych obszarów neuronalnych? Jakie rodzaje struktur w świecie albo ciele własnym stanowią przedmioty reprezentacji? Wszystko to są istotne dla rozwijanego tu projektu pytania empiryczne, do których powrócę pod koniec tego rozdziału (w podrozdziale 4.3).

4.1.4. Wewnętrzne reprezentacje ikoniczne i problem roli funkcjonalnej. S-reprezentacje bez interpretujących podmiotów

Drugie wspomniane wcześniej wyzwanie stojące przed zwolennikiem uznania reprezentacji wewnętrznych za rodzaj reprezentacji ikonicznych – to problem roli funkcjonalnej. Jego naturę można wyrazić za pomocą następujących pytań: czy oprócz zewnętrznych S-reprezentacji uda się wyróżnić S-reprezentacje o charakterze wewnętrznym, to znaczy takie, które mogłyby wchodzić w skład *mechanizmów* (potencjalnie) wyjaśniających zdolności poznawcze? Czy możemy mówić o S-reprezentacjach (strukturach grających rolę S-reprezentacji) nawet wtedy, gdy nie są one interpretowane przez podmiot intencjonalny? Czy możemy nadać sens tezie, że wewnętrzne S-reprezentacje – w przeciwieństwie do receptorów, reprezentacji ukrytych czy też reprezentacji jako przewodników działań – rzeczywiście pełnią funkcję reprezentowania czegoś? Czy wewnętrzne S-reprezentacje spełniają Ramseyowski wymóg opisu zadań?

Problem roli funkcjonalnej ma fundamentalne znaczenie w kontekście celów teoretycznych tej książki. Wyznacza on samo sedno prowadzonych tu poszukiwań funkcjonalnej koncepcji reprezentacji w kognitywistyce, czyli takiej, która odpowiadałaby na pytanie, co to znaczy odgrywać rolę reprezentacji w ramach mechanizmu poznawczego. Przypomnijmy, że zgodnie z opisaną w poprzednim rozdziale metodą Ramseya reprezentacje postulowane w mechanistycznych wyjaśnieniach z zakresu nauk kognitywnych będą „zawdzięczać" swój status reprezentacji temu, że ich rola funkcjonalna może zostać w naturalny i intuicyjnie zrozumiały sposób określona jako rola bycia reprezentacją (rola reprezentowania czegoś). To zaś jest możliwe wtedy, gdy rola tego rodzaju reprezentacji w wystarczającym stopniu

przypomina rolę odgrywaną przez struktury przednaukowo konceptualizowane jako reprezentacje. Odwołując się do tego założenia, chcę bronić tezy, że mogą istnieć wewnętrzne S-reprezentacje, których funkcja rzeczywiście polega na reprezentowaniu czegoś; swój reprezentacyjny status zawdzięczają one zaś temu, że ich rola nietrywialnie przypomina tę pełnioną przez zewnętrzne czy pozamentalne S-reprezentacje. Wewnętrzne S-reprezentacje mogą pod względem funkcjonalnym przypominać pozamentalne, artefaktualne S-reprezentacje, takie jak mapy, diagramy czy makiety.

Ze względu na znaczenie, jakie mają dla mojej argumentacji reprezentacje tego ostatniego rodzaju (zewnętrzne, pozamentalne S-reprezentacje), warto przyjrzeć się im bliżej. Co dokładnie pozamentalne S-reprezentacje „robią" dla swoich użytkowników? Na czym polega ich funkcja? Powiedziałem już, że można ich funkcję uznać ogólnie za *zastępowanie* przedmiotu reprezentacji dla kogoś (por. przypis 4). Warto teraz nieco tę charakterystykę rozszerzyć. Za Swoyerem (1991) proponuję określić rolę pełnioną przez S-reprezentacje jako bycie podstawą dla *rozumowań surogatywnych* (por. też Ramsey 2007: 79). Rozumowanie surogatywne polega na tym, że pewna osoba (1) przeprowadza określone działanie, czysto poznawcze lub praktyczne, na nośniku reprezentacji oraz (2) wyciąga na tej podstawie wnioski dotyczące przedmiotu reprezentacji. Nośnik S-reprezentacji stanowi zatem dla użytkownika „surogat" przedmiotu reprezentacji. W taki właśnie sposób, obserwując strukturę nośnika S-reprezentacji, możemy dowiedzieć się wiele o strukturze przedmiotu reprezentacji. Na przykład analizując układ relacji zachodzących między elementami drzewa genealogicznego, jesteśmy w stanie poprawnie określać relacje pokrewieństwa w danej rodzinie. Rozumowania surogatywne mogą być też przeprowadzane w celu podjęcia praktycznych decyzji dotyczących działań zorientowanych na przedmiot reprezentacji. Poszukując przykładowo najkrótszej drogi łączącej dwa budynki w obrębie pewnego terenu, możemy surogatywnie posłużyć się mapą i sprawdzić którędy przebiega najkrótsza linia między odpowiednimi punktami. Kiedy zaś mamy do czynienia z dynamicznymi S-reprezentacjami, możemy surogatywnie wnioskować o przebiegu procesów zachodzących w przed-

miocie reprezentacji. Na przykład dysponując ruchomą makietą Układu Słonecznego, jesteśmy zdolni przewidzieć relatywne położenie planet w jakimś odstępie czasu. Co więcej, odpowiednio korzystając z takiej dynamicznej S-reprezentacji, możemy wnioskować o tym, jak przedmiot reprezentacji zachowywałby się w pewnych kontrfaktycznych okolicznościach. Dla przykładu zachowanie miniatury samolotu w tunelu aerodynamicznym może stanowić dla inżyniera podstawę do wyciągnięcia wniosków na temat tego, jak docelowa (reprezentowana) konstrukcja zachowywałaby się w analogicznych warunkach. Rozumowanie surogatywne polega więc zawsze na określonym operowaniu[14] nośnikiem reprezentacji w celu wyciągania wniosków, przewidywania zachowań czy też podejmowania praktycznych decyzji dotyczących przedmiotu reprezentacji[15]. Dzięki działaniom wykonywanym na nośniku reprezentacji możemy zrealizować poznawcze lub praktyczne cele dotyczące jej przedmiotu.

Proponuję uznać, że wewnętrzne, mentalne S-reprezentacje mogą pełnić w systemie poznawczym rolę polegającą – w nietrywialnym, wartościowym eksplanacyjnie sensie – na reprezentowaniu czegoś. Są one „mechanicznymi" wersjami zewnętrznych, pozamentalnych S-reprezentacji: pełnią w zasadzie tę samą funkcję, pomimo tego że nie są interpretowane przez pełnoprawny podmiot intencjonalny. Mając na uwadze przedstawioną wyżej charakterystykę funk-

[14] Jak widać, terminu „rozumowanie" nie należy interpretować dosłownie. Nie musi on dotyczyć procesów *stricte* inferencyjnych. Jest on zastosowany na tyle inkluzywnie, że nawet czysto fizyczne ingerowanie w nośnik reprezentacji można uznać za formę surogatywnego „rozumowania". Warunkiem jest to, by taka ingerencja w nośnik była stosowana w celu wyciągnięcia wniosków na temat przedmiotu reprezentacji albo pokierowania działaniami wykonywanymi względem niego.

[15] Nie zapominajmy tu jednak o roli, jaką pełni relacja podobieństwa strukturalnego między nośnikiem a przedmiotem reprezentacji. Przeprowadzanie udanych rozumowań surogatywnych wymaga tego, aby rzeczywiście zachodziło podobieństwo strukturalne między nimi. Jeśli ono nie zachodzi (w wymaganym stopniu czy zakresie), to przeprowadzanie wnioskowań surogatywnych będzie prowadziło do błędnych wniosków, decyzji i predykcji dotyczących przedmiotu reprezentacji. Wykonując rozumowania surogatywne, wykorzystujemy relację podobieństwa między nośnikiem a przedmiotem reprezentacji.

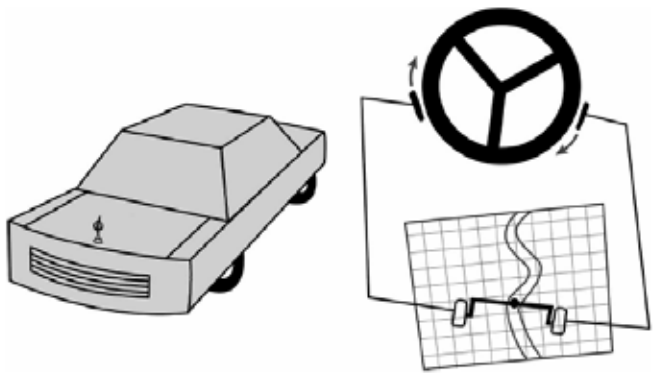
cjonalną zewnętrznych S-reprezentacji, można teraz precyzyjniej wyrazić tę ideę. Otóż proponuję uznać, że wewnętrzne S-reprezentacje stanowią (mogą stanowić) podstawę dla swego rodzaju „zmechanizowanych" rozumowań surogatywnych. W taki właśnie sposób reprezentacje tego rodzaju spełniają wymóg opisu zadań. Jest to diagnoza zbieżna z tym, jak status S-reprezentacji w kognitywistyce ocenia Ramsey (2007: 77–92). Aby uzasadnić przyjmowane tu stanowisko, chcę raz jeszcze powołać się na argumentację przedstawioną przez tego autora. Na tej podstawie rozwinę (w podrozdziale 4.2) pozytywną koncepcję mechanizmów reprezentacyjnych.

Procedura, jaką przyjmuje Ramsey (2007: 80–86, 194–201), argumentując za wartością eksplanacyjną wewnętrznych, czysto mechanicznych S-reprezentacji, jest następująca. Po pierwsze, każe on nam wyobrazić sobie niekontrowersyjny przykład S-reprezentacji o charakterze zewnętrznym i pozamentalnym, to znaczy takiej, która funkcjonuje jako reprezentacja, ponieważ jest w odpowiedni sposób wykorzystywana (interpretowana) przez podmiot intencjonalny. Po drugie, modyfikuje on wyjściowy przykład w taki sposób, że to, co stanowiło w nim S-reprezentację, nadal spełnia maksymalnie podobną rolę funkcjonalną, jednak nie jest wykorzystywane (interpretowane) jako reprezentacja przez żaden podmiot. Po trzecie, Ramsey pokazuje, że w tak zmodyfikowanym przykładzie nadal mamy do czynienia z S-reprezentacją w wartościowym eksplanacyjnie sensie; to znaczy w zmodyfikowanym przypadku określona struktura nadal pełni funkcję, którą możemy w dobrze uzasadniony, nietrywialny sposób scharakteryzować jako umożliwianie czy bycie podstawą rozumowań surogatywnych. Przyjrzyjmy się dwóm przykładom zastosowania takiej strategii przez Ramseya.

Przypadek 1: samochód Cumminsa

Wyobraźmy sobie samochód poruszający się torem w kształcie litery „S", którego granice są wyznaczane przez wysoki mur. Samochód nie odbija się od krawędzi do krawędzi toru, lecz jedzie środkiem, zakręcając w odpowiednich momentach, pod odpowiednim kątem i w odpowiednim kierunku. Chcemy wyjaśnić, jak to się dzieje, że samochód pokonuje tor w taki sposób. Odkrywamy, że wewnątrz

znajduje się kierowca. Okazuje się jednak, że okna samochodu są nieprzezroczyste, a zatem kierowca nie widzi toru, po którym się porusza. Zamiast tego, prowadzi on samochód wykorzystując mapę odzwierciedlającą fizyczny kształt toru. Jest to mapa elektroniczna, w której samochód oznaczono za pomocą czerwonego punktu. Punkt ten porusza się po mapie równocześnie z tym, jak samochód porusza się po torze, w taki sposób, że położenie punktu na mapie odpowiada każdorazowo położeniu samochodu w obrębie toru. Kierowca śledzi położenie punktu na mapie i na tej podstawie kieruje samochodem. Wykorzystuje on zatem mapę jako S-reprezentacyjny surogat toru.



Rysunek 3. Opisany pierwotnie przez Roberta Cumminsa, w pełni mechaniczny samochód, który przemierza tor, wykorzystując wewnętrzną mapę (S-reprezentację). Źródło: Ramsey 2007: 199

Zmodyfikujemy teraz ten przykład. Patrząc z zewnątrz, przypadek ten jest nieodróżnialny od tego opisanego powyżej: widzimy sprawnie poruszający się po torze samochód. Kiedy jednak zaglądamy do środka, okazuje się, że pojazd nie jest kierowany przez człowieka. Zamiast siedzącego w środku kierowcy odkrywamy, że we wnętrzu samochodu znajduje się niewielka tablica, w której wydrążono rowek (por. rysunek 3). Rowek ten ma kształt litery „S", odpowiadają-

cy kształtowi toru, po którym porusza się samochód. Do rowka podłączono ster poruszający się wewnątrz rowka równocześnie z tym, jak pojazd porusza się w obrębie toru, w taki sposób, że ster każdorazowo znajduje się w miejscu rowka, które odpowiada umiejscowieniu samochodu na torze. Ster jest podłączony do kierownicy tak, że kierunek jej obrotu – a przez to kierunek poruszania się samochodu – systematycznie odpowiada zmianom orientacji steru. Ster „przetwarza" kształt rowka na takie ruchy kierownicą, które zapewniają samochodowi skuteczne poruszanie się po torze. Biorąc pod uwagę, że taki pozbawiony kierowcy samochód został pierwotnie opisany nie przez samego Ramseya (on adaptuje go po prostu do własnych celów), lecz Roberta Cumminsa (1996), nazwijmy ten pojazd skrótowo „samochodem Cumminsa".

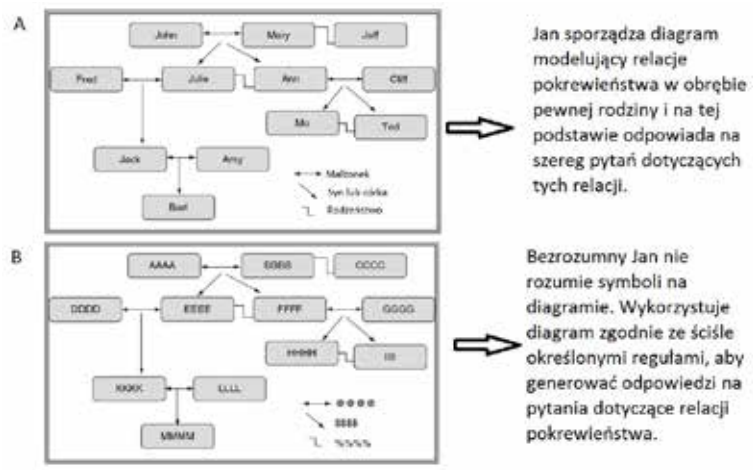
Czy samochód Cumminsa stanowi system, którego działanie powinniśmy wyjaśniać w kategoriach reprezentacyjnych? Czy mechanizm odpowiadający za sterowanie jego ruchem wykorzystuje reprezentację toru? Czy wewnętrzny rowek, do którego jest podłączony ster, rzeczywiście stanowi funkcjonalny odpowiednik mapy, jaką posługiwał się kierowca w naszym wyjściowym przykładzie? Wydaje się, że odpowiedź w każdym przypadku powinna być twierdząca (Ramsey 2007: 198–199). Sposób funkcjonowania rowka przypomina ten, w jaki funkcjonowała mapa w samochodzie kierowanym przez człowieka. Samochód Cumminsa osiąga sukces nawigacyjny dzięki temu, że wykorzystuje on strukturalne (przestrzenne) podobieństwo zachodzące między torem a rowkiem. Możemy powiedzieć, że samochód wykorzystuje wewnętrzną strukturę (rowek) zamiast wchodzenia w interakcje z określonym elementem środowiska zewnętrznego (ścianami toru). Rowek stanowi wewnętrzny „surogat" toru. Co więcej, satysfakcjonujące wyjaśnienie sukcesu nawigacyjnego pojazdu – to znaczy wskazanie mechanizmu odpowiedzialnego za ten sukces – wymaga powołania się na fakt, że rowek pełni funkcję wewnętrznej mapy. Jeśli nie spojrzymy na działanie samochodu Cumminsa w ten sposób, nie zrozumiemy tego, jak to się dzieje, że porusza się on po torze tak skutecznie. W szczególności musimy powołać się na fakt, że strukturalno-przestrzenne własności mapy odzwierciedlają strukturalno-przestrzenne własności toru. Co

więcej, ewentualne zaburzenie podobieństwa między rowkiem a torem mogłoby wyjaśniać sytuacje, w których samochód Cumminsa nie pokonuje toru skutecznie. Mamy tu zatem do czynienia z przypadkiem, w którym rowek *reprezentuje* tor na podstawie strukturalnego podobieństwa. Samochód Cumminsa to przykład niezwykle prostego systemu fizycznego posługującego się wewnętrzną, czysto mechaniczną S-reprezentacją. Posługuje się on mapą (S-reprezentacją) toru, mimo że nie ma w nim podmiotu, który mógłby ową mapę (dosłownie) interpretować.

Przypadek 2: Bezrozumny Jan

Wyobraźmy sobie mężczyznę – nazwijmy go „Janem" – który ma za zadanie odpowiedzieć na serię pytań dotyczących relacji pokrewieństwa zachodzących między poszczególnymi członkami pewnej rodziny. Rodzina ta jest jednak liczna, a pamięć Jana niedoskonała, w związku z czym nie jest on w stanie zapamiętać wszystkich relacji, których dotyczą zadawane mu pytania. Aby ułatwić sobie zadanie, Jan postanawia stworzyć diagram odzwierciedlający strukturę pokrewieństwa w rodzinie (por. rysunek 4-A). Diagram ten zawiera imiona członków rodziny, które są połączone różnego rodzaju strzałkami. Poszczególne rodzaje strzałek (na przykład pojedyncza, podwójna, z przerywaną linią i tak dalej) mają odpowiadać poszczególnym rodzajom relacji pokrewieństwa (bycie małżonkiem, bycie rodzeństwem, bycie synem/córką i tak dalej). Struktura tak powstałego diagramu odpowiada strukturze relacji pokrewieństwa zachodzących między członkami rodziny. Jeśli jakaś relacja pokrewieństwa zachodzi między dwoma członkami rodziny, to na diagramie imiona tych osób są połączone odpowiedniego rodzaju strzałką. Spoglądając na diagram, Jan nie tylko „widzi" zachodzące w rodzinie relacje, ale też może wyciągnąć nowe wnioski na temat relacji pokrewieństwa, z których zachodzenia nie zdawał sobie wcześniej sprawy. Dzięki użyciu diagramu Jan jest w stanie udzielać poprawnych odpowiedzi na zadawane mu pytania. Kompensuje on tym samym ograniczenia własnej pamięci, świadomie wykorzystując S-reprezentację interesującej go domeny (przeprowadzając surogatywne rozumowania na jej temat).

Zastąpmy teraz Jana osobą, która stoi przed tym samym diagramem, lecz nie rozumie przedstawionych na nim symboli (por. rysunek 4-B). Nie potrafi ona odczytać zapisanych na diagramie imion członków rodziny, nie rozumie też towarzyszącej diagramowi notki określającej, którym rodzajom relacji pokrewieństwa odpowiadają poszczególne rodzaje strzałek. Nazwijmy tę osobę „Bezrozumnym Janem". Bezrozumny Jan nie wie nawet, że znajdujący się przed nim diagram stanowi reprezentację relacji pokrewieństwa w pewnej rodzinie. Mimo to, między strukturą diagramu a strukturą tej rodziny nadal zachodzi podobieństwo. Wyobraźmy sobie, że Bezrozumny Jan musi odpowiedzieć na szereg pytań dotyczących relacji pokrewieństwa zachodzących w reprezentowanej przez diagram rodzinie. Są to te same pytania, na które wcześniej odpowiadał Jan. W przeciwieństwie do Jana, Bezrozumny Jan nie jest w stanie zrozumieć znaczenia zadawanych mu pytań, tak samo jak nie rozumie elementów diagramu. Zamiast tego dysponuje on całym szeregiem czysto „syntaktycznych" reguł określających, co należy zrobić (jakiej odpowiedzi należy udzielić), jeśli na diagramie dwa napisy o określonym kształcie fizycznym są połączone strzałką o określonym kształcie fizycznym. Reguły te stworzono w sposób respektujący odpowiedniość między elementami diagramu i rodziny (to znaczy respektując to, które imiona i strzałki na diagramie odpowiadają którym osobom oraz relacjom pokrewieństwa); pozwalają one w ten sposób na „przetworzenie" elementów diagramu na zdania opisujące relacje pokrewieństwa w rodzinie. Ilekroć zostaje zadane pytanie, Bezrozumny Jan spogląda na diagram i na podstawie odpowiedniej reguły udziela odpowiedzi. Choć nie ma on pojęcia, co robi i jak to robi, to – dzięki strukturalnemu podobieństwu zachodzącemu między diagramem a układem pokrewieństwa w rodzinie – udaje mu się udzielić szeregu poprawnych odpowiedzi. Dla zewnętrznego obserwatora sposób jego działania może być nieodróżnialny od tego, jak na te same pytania odpowiadał „rozumny" Jan z wyjściowego przykładu.



Rysunek 4. Diagram relacji rodzinnych jako S-reprezentacja. A. Diagram taki, jakim go widzi Jan. B. Diagram taki, jakim go widzi Bezrozumny Jan. Źródło: Ramsey 2007: 81, 84

Czy Bezrozumny Jak korzysta jednak z reprezentacji? Czy postępuje on w sposób na tyle podobny do tego, jak postępował („rozumny") Jan, by można było powiedzieć, że mamy tu do czynienia z wykorzystywaniem reprezentacji? Raz jeszcze wydaje się, że odpowiedzi powinny być twierdzące (Ramsey 2007: 84–85). Rola funkcjonalna diagramu w przykładzie z Janem jest w zasadniczy sposób zbliżona do roli pełnionej przez ten sam diagram w przykładzie z Bezrozumnym Janem. Aby odpowiedzieć na pytanie o to, jak Janowi udaje się udzielać poprawnych odpowiedzi na zadawane mu pytania, musimy odwołać się do faktu, że polega on w swoich staraniach na strukturalnym podobieństwie między diagramem a układem pokrewieństwa w rodzinie. To samo dotyczy jednak Bezrozumnego Jana. Także on wykorzystuje relację podobieństwa zachodzącą między diagramem a reprezentowaną rodziną. Gdyby taka relacja nie zachodziła, diagram nie pozwoliłby Bezrozumnemu Janowi poprawnie odpowiadać na zadawane mu pytania. Jan i jego bezrozumny odpowiednik

wykonują ten sam rodzaj rozumowania surogatywnego, z tą różnicą, że jeden robi to w sposób „rozumiejący" a drugi – w sposób pozbawiony „rozumienia". Obydwaj korzystają, choć na nieco różne sposoby, z S-reprezentacji określonej domeny. Gdybyśmy nie przyjęli, że także Bezrozumny Jan wykorzystuje reprezentację relacji pokrewieństwa w rodzinie, to nie moglibyśmy zrozumieć, jak udaje mu się odnieść sukces przy odpowiadaniu na pytania. Ramsey ujmuje to następująco: „zagadka, jak Bezrozumny Jan odkrył relację pokrewieństwa, zostałaby zastąpiona zagadką, jak Bezrozumny Jan odkrył relację pokrewieństwa, bawiąc się elementami diagramu" (Ramsey 2007: 84). Sukces Bezrozumnego Jana w zasadniczy sposób zależy od tego, że między diagramem a układem relacji pokrewieństwa zachodzi strukturalne podobieństwo. Jeśli chcemy ten sukces wyjaśnić, musimy powołać się na fakt, że Bezrozumny Jan korzysta (choć bezwiednie) z owego podobieństwa. Istnieje zatem eksplanacyjny zysk z interpretowania tego przypadku z (S-)reprezentacyjnej perspektywy, zysk, który u swoich podstaw ma rolę funkcjonalną, jaką dla Bezrozumnego Jana spełnia diagram relacji rodzinnych.

Można tu oponować: Bezrozumny Jan jest przecież intencjonalnym podmiotem. Tymczasem powinna interesować nas możliwość (zasadnego) wykorzystania pojęcia S-reprezentacji nawet wtedy, gdy reprezentacje te nie są przez kogoś (pewien podmiot) interpretowane. Krytyka ta będzie jednak nieskuteczna. Kiedy przyjrzymy się bliżej zdolnościom Bezrozumnego Jana, szybko okazuje się, iż jest on na tyle bezrozumny, że w istocie przypomina działający czysto „syntaktycznie" system obliczeniowy. Bezrozumnego Jana możemy bez problemu w naszym przykładzie zastąpić – otrzymując te same konkluzje – sztucznym, mechanicznym układem realizującym zestaw algorytmów określających reguły udzielania odpowiedzi na zadawane pytania (Ramsey 2007: 84)[16].

[16] Bezrozumny Jan przypomina pod tym względem osobę zamieszkującą Searle'owski chiński pokój (Searle 1980). Zaznaczmy jednak wyraźnie, że przykład z Bezrozumnym Janem nie ma być rozumiany jako argument za tezą, której swój eksperyment myślowy *przeciwstawił* John Searle. Nie chodzi więc o to, że Bezrozumny Jan zaczyna *rozumieć* znaczenie czy semantykę symboli przedstawionych na diagramie; chodzi o to, że aby wyjaśnić odniesiony przez

Jeśli zgodzimy się na zaproponowane powyżej, (S-)reprezentacyjne wyjaśnienia działania samochodu Cumminsa oraz Bezrozumnego Jana, to wydaje się, że problem roli funkcjonalnej powinniśmy uznać za rozwiązany. Obydwa te przykłady w jasny sposób pokazują, że coś może odgrywać rolę S-reprezentacji, nawet jeśli nie podlega ono interpretacji przez pełnoprawny podmiot intencjonalny. Są to bowiem przypadki, gdy pewien fizyczny system koordynuje swoje działanie (poznawcze lub praktyczne) względem pewnego przedmiotu, a udaje mu się to dzięki temu, że odpowiednio wykorzystuje on strukturalne podobieństwo zachodzące między tym przedmiotem a pewną wewnętrzną strukturą (komponentem mechanizmu). Ta wewnętrzna struktura pełni funkcję nośnika reprezentacji, a dokładniej – nośnika S-reprezentacji. Możemy zasadnie powiedzieć, że samochód Cumminsa i Bezrozumny Jan stosują pewną wersję rozumowań surogatywnych. Są wyposażeni w mechanizmy realizujące pewne funkcje związane z *przedmiotem* reprezentacji dzięki odpowiedniemu wykorzystaniu jej wewnętrznych *nośników*. Nasze pojęcie reprezentacji w kognitywistyce jako czysto mechanicznych S-reprezentacji spełnia wymóg opisu zadań.

Samochód Cumminsa i Bezrozumny Jan to oczywiście jedynie poglądowe, czysto „zabawkowe" przykłady. Czy kognitywiści powołują się jednak na wewnętrzne struktury, które funkcjonują w podobny sposób? Czy (mechanistyczne) wyjaśnienia w kognitywistyce odwołują się do wewnętrznych struktur systemu poznawczego, których sposób funkcjonowania przypomina to, jak funkcjonuje wewnętrzny rowek w samochodzie Cumminsa albo diagram w przykładzie z Bezrozumnym Janem? Sądzę, że tak. Konkretne przykłady takiej praktyki ekplanacyjnej wskażę w podrozdziale 4.3. Póki co, bardziej zasadniczy cel tej pracy nie został jednak w pełni zrealizowany. Przebieg dotychczasowych rozważań pozwala sądzić, że na pojęciu S-reprezentacji można oprzeć koncepcję mechanizmów re-

Bezrozumnego Jana sukces, musimy powołać się na fakt, iż (nieświadomie i „bezrozumnie") polega on na S-reprezentacji domeny, której dotyczą zadawane mu pytania (por. Ramsey 2007: 89–90).

prezentacyjnych. Jednak do tej pory nie dostarczyłem takiej koncepcji. Czym są zatem mechanizmy reprezentacyjne?

4.2. Mechanizmy reprezentacyjne jako mechanizmy wyposażone w konsumowane modele

4.2.1. Mechanizmy, S-reprezentacje i konsumowane modele

Usytuujmy dotychczasowy wywód w kontekście omówionej w poprzednim rozdziale metody Ramseya (por. sekcja 3.2.2). W pierwotnej wersji aplikacja tej metody kończy się wykonaniem kroku (c), w którym to dokonuje się oceny danego pojęcia reprezentacji pod kątem spełniania przez nie wymogu opisu zadań. Należy jednak przypomnieć, że w celu zaadaptowania metody Ramseya do wymogów realizowanego w tej pracy projektu zdecydowałem się uzupełnić ją jeszcze jednym, czwartym krokiem. Warunkiem możliwości wykonania tego kroku jest to, by badane przez nas pojęcie reprezentacji spełniało wymóg opisu zadań, pozytywnie przechodząc ewaluację wykonaną w kroku (c). Jeśli tak jest, to ostatnie zadanie, wykonywane właśnie w kroku (d), polega na wyrażeniu czy konceptualizacji roli funkcjonalnej pełnionej przez desygnaty tego pojęcia reprezentacji w kategoriach mechanistycznych. Powinniśmy zatem przyjąć, że interesujące nas reprezentacje mogą pełnić rolę reprezentacji w mechanizmach, stanowiąc ich *aktywne komponenty*. Następnie powinniśmy odpowiedzieć na pytanie: co takiego w funkcjonowaniu tych komponentów czyni je reprezentacjami? Inaczej: co jest takiego w roli (operacji), jakie pełnią one względem innych komponentów mechanizmu, że powinniśmy rozpatrywać tę rolę jako reprezentowanie czegoś? Jeśli zaś wiemy już, co znaczy dla komponentu mechanizmu bycie reprezentacją, to powinniśmy być także zdolni do udzielenia odpowiedzi na pytanie o to, co czyni określone *mechanizmy* – reprezentacyjnymi. W taki właśnie sposób zastosowanie metody Ramseya ma umożliwić sformułowanie koncepcji mechanizmów reprezentacyjnych.

Wniosek płynący z prowadzonych w poprzednim podrozdziale rozważań można podsumować za pomocą twierdzenia: S-reprezentacje (pojęcie S-reprezentacji) spełniają wymóg opisu zadań. Wykonanie kroku (c) metody Ramseya w ich przypadku przynosi wynik pozytywny. Skoro tak, to zgodnie z przywołaną wyżej procedurą należy wykonać kolejny krok i zadać sobie pytanie: dzięki czemu S-reprezentacje spełniają wymóg opisu zadań? Przyjmując, że S-reprezentacje stanowią aktywne komponenty mechanizmów, pytanie brzmi: co takiego w tym, co „robią" one dla innych komponentów mechanizmu, czyni je reprezentacjami? Na czym polega „położenie" funkcjonalne S-reprezentacji w kontekście szerszego mechanizmu? Kiedy odpowiemy sobie na te pytania, będziemy dysponować teorią mechanizmów reprezentacyjnych.

Twierdzę, że S-reprezentacja odgrywa w ramach mechanizmu rolę *konsumowanego modelu* pewnej domeny. Pominę na razie zagadnienie konsumentów/konsumpcji i skupię się na wykorzystaniu terminu „model" w tym sformułowaniu. Mówiąc o modelach, przyświeca mi cel nie tyle teoretyczny, ile raczej terminologiczny: chcę ten termin stosować zamiast „S-reprezentacji". Model rozumiem po prostu jako reprezentację, która odzwierciedla strukturalnie swój przedmiot. Dotyczy to zarówno modeli przyjmujących postać publicznych artefaktów kulturowych (w tym modeli naukowych), jak i właśnie modeli stanowiących komponenty wewnętrznych mechanizmów poznawczych. Te ostatnie będę tu nazywać „modelami wewnętrznymi" lub „mentalnymi"[17]. Choć mówienie o „modelach"

[17] Wprowadzenie terminu „model mentalny" wydaje się w naturalny sposób sytuować te rozważania w kontekście teorii modeli mentalnych Philipa Johnsona-Lairda (1983; por. także Piłat 1999, gdzie autor rozwija koncepcję Johnsona-Lairda i aplikuje ją przy rozstrzyganiu problemów filozoficznych). Należy jednak zaznaczyć, że związki między rozwijaną tu koncepcją a teorią Johnsona-Lairda są ograniczone. Teoria Johnsona-Lairda jest co prawda jak najbardziej zbieżna z naszą teorią mechanizmów reprezentacyjnych, jednak ta ostatnia nie powinna być traktowana jako rozwinięcie czy uzupełnienie tej pierwszej. Johnson-Laird postuluje istnienie konkretnego rodzaju modeli mentalnych, mianowicie swoistych wewnętrznych „diagramów" wykorzystywanych przy przeprowadzaniu rozumowań (na przykład podczas rozwiązywania sylogizmów). Tymczasem w ramach swej pracy chcę odpowiedzieć na bardziej zasadnicze pytanie o to, co

zamiast o „S-reprezentacjach" jest zabiegiem dotyczącym nomenklatury, to jednak wydaje się, że w ten sposób udaje się uwypuklić ważny aspekt prezentowanego tu podejścia do natury reprezentacji. Termin „S-reprezentacja" kładzie nacisk na relację zachodzącą między nośnikiem a przedmiotem reprezentacji, czyli relację podobieństwa strukturalnego. Mówiąc zamiast tego o modelach, chcę podkreślić fundamentalne znaczenie *użycia* pewnej struktury jako reprezentacji. Modele są zawsze modelami *dla* kogoś (lub, jak się okaże, dla czegoś), stosowanymi w jakimś celu. Sama idea modelu wydaje się w nieunikniony sposób powiązana z ideą jego użytkownika (por. Giere 2004; 2010). Jak stwierdza Kenneth Craik – pierwszy autor[18], który posłużył się terminem „model" w kontekście rozważań nad naturą reprezentacji mentalnych:

> Jeśli organizm posiada w głowie „pomniejszony model" zewnętrznej rzeczywistości i swoich własnych działań, to jest zdolny wypróbowywać różne alternatywny [alternatywne działania – P. G.], ustalać, która z nich jest najlepsza, reagować na przyszłe sytuacje, wykorzystywać wiedzę o zdarzeniach uprzednich, aby sprostać wymaganiom teraźniejszości i przyszłości, oraz, mówiąc ogólnie, reagować na powstające zagrożenia w sposób bardziej kompletny, bezpieczny i umiejętny (Craik 1943: 61).

Wedle Craika – dzięki dysponowaniu wewnętrznym modelem pewien system (mechanizm) może osiągać określone cele praktyczne. Polega on na modelu, aby, na przykład, przewidywać przyszłe zdarzenia i na tej podstawie decydować o tym, które działania są sto-

---

to w ogóle znaczy, że pewien system fizyczny posługuje się wewnętrznym modelem, bez względu na to, czym dokładnie miałby taki model być i jakie miałby posiadać konkretne zastosowania w systemie. Innymi słowy, Johnson-Laird w swojej teorii porusza się na poziomie przedmiotowym (wyjaśnia on określone zjawiska, postulując wewnętrzne modele o określonych własnościach), natomiast koncepcja rozwijana w tym rozdziale dotyczy poziomu metaprzedmiotowego (pragnę określić, na czym w ogóle polega wyjaśnienie pewnego zjawiska za pomocą reprezentacji mentalnych rozumianych jako modele).

[18] Taką pionierską rolę Craikowi przypisuje Johnson-Laird (2005).

sowne w danej sytuacji. Decydując się na termin „model", pragnę podkreślić praktyczne znaczenie tego rodzaju reprezentacji.

Posiadanie owego praktycznego znaczenia czy wartości przez model wewnętrzny jest możliwe dzięki relacji strukturalnego podobieństwa zachodzącej między nośnikiem (samym modelem) a przedmiotem reprezentacji. Jednakże samo strukturalne podobieństwo nie wystarcza do tego, aby coś odgrywało rolę modelu. Na czym więc polega *pełnienie funkcji* modelu w ramach mechanizmu? Właśnie tu pojawia się pojęcie konsumenta reprezentacji. Proponuję uznać, że wewnętrzne (mentalne) S-reprezentacje pełnią funkcję modeli, o ile posiadają w mechanizmie *konsumenta*. Konsumentem modelu jest ten komponent mechanizmu, który wykorzystuje ów model – to znaczy wykorzystuje komponent stanowiący nośnik wewnętrznej S-reprezentacji – przy realizowaniu własnych funkcji (operacji). Wewnętrzne modele są modelami dla swoich konsumentów, ponieważ można o nich powiedzieć, że „przewodzą działaniami" tych ostatnich, w sensie zbliżonym do teorii Andersona i Rosenberga (2008).

Kategoria konsumenta zostanie jeszcze rozjaśniona w toku prowadzonego tu wywodu. Najpierw spróbuję jednak na podstawie powyższych twierdzeń sformułować teorię mechanizmów reprezentacyjnych. Otóż twierdzę, że mechanizm reprezentacyjny to *mechanizm wyposażony w konsumowany model* (dalej: MKM). MKM to zaś mechanizm spełniający cztery następujące warunki:

a) Występowanie nośnika reprezentacji i przedmiotu reprezentacji powiązanych relacją podobieństwa strukturalnego

MKM zawiera komponent, którego poprawne funkcjonowanie zależy od tego, czy między tym komponentem a czymś zewnętrznym względem niego zachodzi relacja podobieństwa strukturalnego. Komponent taki stanowi *nośnik reprezentacji*. To, do czego nośnik jest czy powinien być – w zgodzie ze swoją funkcją lub realizowanym w mechanizmie działaniem – strukturalnie podobny, stanowi *przedmiot reprezentacji*[19].

[19] Inaczej mówiąc, przedmiot reprezentacji to *warunki jej poprawności*; to struktura, do której nośnik byłby podobny, gdyby stanowił w danej sytuacji

b) Występowanie konsumenta reprezentacji i jego zależność od nośnika reprezentacji

Sposób funkcjonowania co najmniej jednego spośród pozostałych komponentów MKM jest w regularny czy systematyczny sposób zależny przyczynowo od stanu, w jakim znajduje się nośnik reprezentacji. Taki komponent mechanizmu związany przyczynowo z nośnikiem to *konsument reprezentacji*[20]. Przyczynowa relacja z nośnikiem reprezentacji jest konieczna, ale niewystarczająca do bycia konsumentem (dalsze warunki nakładane na bycie konsumentem reprezentacji są wymienione w kolejnych punktach).

c) Konieczność dostosowania konsumenta do przedmiotu reprezentacji, przy jednoczesnym braku bezpośredniej interakcji między nimi

Poprawne działanie konsumenta reprezentacji w MKM zależy od tego, w jakim stanie znajduje (znajdzie, znalazł) się przedmiot reprezentacji. Aby poprawnie realizować własną operację (funkcję w mechanizmie), konsument musi *dostosowywać* sposób swojego funkcjonowania do przedmiotu reprezentacji. Jednak nie występuje bezpośrednia przyczynowa interakcja między konsumentem a przedmiotem reprezentacji.

d) Nośnik reprezentacji jako pośrednik między przedmiotem a konsumentem reprezentacji

Konsument jest nie tylko przyczynowo, ale również funkcjonalnie powiązany z nośnikiem reprezentacji. Konsument działa zatem poprawnie (poprawnie realizuje swoją funkcję) w ramach mechanizmu, pod warunkiem że zachodzi strukturalne podobieństwo między nośnikiem a przedmiotem reprezentacji. Jeśli takie podobieństwo nie zachodzi, konsument nie realizuje poprawnie swojej funkcji w mechanizmie. Poprawne działanie konsumenta zależy więc systematycznie od poprawnego działania nośnika reprezentacji, a popraw-

reprezentację poprawną. Do kwestii tej powrócimy w sekcji 4.2.2, w której odniosę się, między innymi, do zagadnienia treści modeli mentalnych.

[20] Dopuszczam możliwość, że jeden MKM ma więcej niż jeden komponent będący konsumentem.

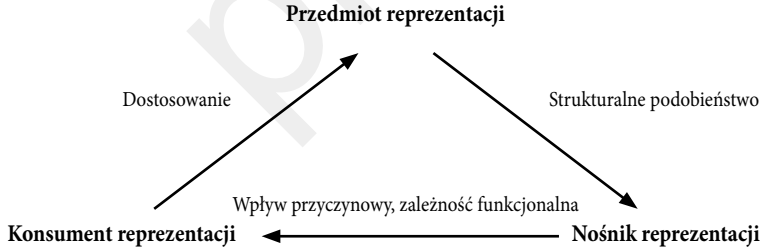
ne działanie nośnika zależy od tego, czy między nim a przedmiotem reprezentacji zachodzi relacja podobieństwa strukturalnego. Nośnik pośredniczy między konsumentem a przedmiotem reprezentacji.

Aby zobrazować tę koncepcję, warto pokazać, w jaki sposób cztery wymienione warunki są spełniane przez jakiś przykładowy MKM. Wykorzystajmy w tym celu opisany w sekcji 4.1.4 mechanizm odpowiadający za poruszanie się samochodu Cumminsa. Mechanizm ten może być sklasyfikowany jako MKM, ponieważ spełnia powyższe cztery warunki w następujący sposób:

Ad a) Nośnikiem reprezentacji jest wewnętrzna mapa toru, czyli tablica, w której wydrążono rowek odzwierciedlający kształt toru. Przedmiotem reprezentacji jest tor pokonywany przez samochód. Poprawne funkcjonowanie tablicy z rowkiem zależy od tego, czy zachodzi podobieństwo strukturalne (przestrzenne) między kształtem rowka a kształtem toru.
Ad b) Konsumentem reprezentacji jest w samochodzie Cumminsa komponent zawiadujący poruszaniem się pojazdu, czyli ster (podłączony do kierownicy). Ster jest w systematyczny sposób powiązany przyczynowo z tablicą, w której wyryto rowek (czyli – z nośnikiem reprezentacji). Ta zależność przyczynowa cechuje się tym, że kierunek i kąt zwracania się steru są systematycznie zależne od kształtu rowka wydrążonego w tablicy.
Ad c) Funkcją steru (konsumenta reprezentacji) w mechanizmie jest odpowiednie sterowanie ruchem pojazdu. To, czy ster poprawnie kieruje samochodem, zależy jednak od kształtu toru (przedmiotu reprezentacji). Gdyby tor miał kształt odwróconej litery „S", to ster – aby poprawnie realizować swoją funkcję – musiałby działać w inny sposób (obracać się w innym kierunku lub pod innym kątem) niż w sytuacji, gdy tor ma kształt litery „S". Ster musi zatem dostosowywać swoje działanie do kształtu toru. Nie wchodzi on jednak w bezpośrednie interakcje przyczynowe z torem.
Ad d) Mechanizm odpowiedzialny za ruch pojazdu jest zorganizowany w taki sposób, że ster (konsument reprezentacji)

poprawnie pełni swoją funkcję – to znaczy umożliwia samochodowi pokonanie toru – pod warunkiem że rowek wyryty w tablicy (nośnik reprezentacji) jest strukturalnie podobny do toru (przedmiotu reprezentacji). Jeśli kształt rowka nie odzwierciedla kształtu toru, ster nie będzie odpowiednio zawiadywał poruszaniem się samochodu.

Twierdzę, że mechanizm sterujący ruchem samochodu Cumminsa zawdzięcza status mechanizmu reprezentacyjnego dzięki temu, że spełnia on warunki (a)–(d). Powiedzieć, że mechanizm ten wykorzystuje reprezentację toru, to inaczej powiedzieć, że jeden z jego komponentów wykorzystuje inny komponent jako model (S-reprezentację) określonej domeny. Ster (konsument reprezentacji) wykorzystuje tablicę z wydrążonym rowkiem (nośnik reprezentacji) zamiast samego toru, po to, aby orientować samochód względem toru (przedmiotu reprezentacji). W ten sposób rowek wydrążony w tablicy *reprezentuje* kształt toru *dla* steru podłączonego do kierownicy. Ster wchodzi w interakcje z wewnętrzną mapą jako „surogatem" toru. Wnioski te można zgeneralizować. Sądzę, że kiedykolwiek me-



Rysunek 5. Schemat przedstawiający wzajemną zależność trzech elementów, które nadają mechanizmowi status MKM. Nośnik reprezentacji powinien zgodnie ze swoim przeznaczeniem znajdować się w relacji podobieństwa strukturalnego do przedmiotu reprezentacji. Kiedy tak rzeczywiście jest, nośnik powinien wpływać na konsumenta w taki sposób, że działanie tego ostatniego będzie dostosowane do stanu, w jakim znajduje (znajdzie, znalazł, znalazłby) się przedmiot reprezentacji

chanizm spełnia warunki (a)–(d), można w uzasadniony sposób powiedzieć, że jest on wyposażony w wewnętrzny model. Bycie MKM sprawia zaś, że dany mechanizm ma charakter reprezentacyjny.

Zwróćmy uwagę na fakt, że koncepcja mechanizmów reprezentacyjnych jako MKM dobrze wpisuje się w Peirce'owski triadyczny sposób rozumienia reprezentacji, który został omówiony w sekcji 4.1.1. Opisany tam układ „nośnik–przedmiot–interpretacja" przyjmuje w MKM postać układu „nośnik–przedmiot–konsument" (por. rysunek 5). Pamiętajmy zarazem, że zgodnie z zaproponowaną w sekcji 4.1.1 rekonstrukcją poszczególne elementy Peirce'owskiej triady muszą być ze sobą powiązane w taki sposób, aby zostały spełnione dwa warunki. Zobaczmy, czy warunki te są spełniane w ramach MKM.

Zgodnie z pierwszym z wymienionych warunków:

> **(W1)** Nośnik reprezentacji jest interpretowany jako reprezentacja przedmiotu reprezentacji. Dzięki temu może on pełnić funkcję polegającą na *zastępowaniu* przedmiotu reprezentacji dla kogoś/czegoś.

Kiedy mowa o reprezentacjach pozamentalnych, warunek ten dotyczy zależności między nośnikiem reprezentacji a interpretacją. W kontekście modeli wewnętrznych warunek ten powinniśmy zaś rozumieć jako dotyczący relacji między nośnikiem reprezentacji a *konsumentem*. W przyjmowanej tu przeze mnie koncepcji konsument stanowi bowiem komponent mechanizmu, który jest funkcjonalnym odpowiednikiem osoby dokonującej interpretacji. Tym, czym dla mapy lub modelu naukowego jest interpretująca osoba, tym dla modelu mentalnego będzie jego konsument. Tak jak osoba wykorzystuje zewnętrzną mapę, aby kierować własnymi działaniami w złożonym otoczeniu, tak komponent mechanizmu (konsument) może wykorzystywać wewnętrzną „mapę", aby kierować działaniami tego systemu; tak jak można powiedzieć, że zewnętrzna mapa zastępuje teren jej użytkownikowi, tak można powiedzieć, że wewnętrzna mapa zastępuje teren konsumentowi. W tym drugim przypadku pełnienie roli reprezentacyjnego „zastępstwa" czy surogatu opiera

się na fakcie, że nośnik reprezentacji „kompensuje" konsumentowi fakt, iż ten nie znajduje się w bezpośrednim kontakcie przyczynowym z przedmiotem reprezentacji. Należy bowiem podkreślić: konsument pełni w mechanizmie rolę funkcjonalną, której poprawne realizowanie wymaga dostosowania do pewnego zewnętrznego stanu rzeczy, czyli przedmiotu reprezentacji. Jednak w MKM dostosowanie to nie jest możliwe dzięki nawiązaniu przez konsumenta bezpośrednich interakcji z przedmiotem reprezentacji. W takiej właśnie sytuacji konsument wykorzystuje model przedmiotu reprezentacji *zamiast* wchodzenia z nim w interakcje. Inaczej mówiąc, przeprowadza on na nośniku (czysto „mechanicznie" pojęte) rozumowania surogatywne.

Występujące w prezentowanej tu teorii pojęcie konsumenta reprezentacji zostało rzecz jasna zaczerpnięte z prac Ruth Millikan (1984; 2002). Zachodzą jednak dwie ważne różnice między sposobem rozumienia konsumenta przez Millikan a tym, jak rozumie się go w bronionej tu przeze mnie koncepcji. Pierwsza różnica dotyczy *celów teoretycznych* związanych ze wprowadzeniem pojęcia konsumenta do teorii. Millikan (1984; 2002) jest zainteresowana przede wszystkim problemem treści stanów mentalnych. Mówiąc bardzo skrótowo, wedle tej autorki treść, czyli inaczej warunki poprawności (prawdziwości) reprezentacji, jest wyznaczana przez warunki poprawnego funkcjonowania konsumenta. Konsument i pełniona przez niego funkcja determinują treść reprezentacji w tym sensie, że treścią reprezentacji są warunki, które musiałyby zostać spełnione, aby konsument (biorąc pod uwagę jego funkcję) w danej sytuacji odniósł sukces. Z kolei w proponowanej tutaj koncepcji mechanizmów reprezentacyjnych skupiam się na pytaniu o to, co to znaczy funkcjonować jako reprezentacja w mechanizmie. Kategorię konsumenta wprowadzam właśnie w celu udzielenia odpowiedzi na to właśnie pytanie. Twierdzę, że żaden komponent mechanizmu nie realizuje operacji (funkcji) polegającej na reprezentowaniu czegoś, o ile nie posiada on swoich konsumentów. Mówiąc hasłowo, bez konsumentów reprezentacji nie ma struktur pełniących funkcję jej nośników. Komponent mechanizmu staje się reprezentacją dopiero pod warunkiem, że jest on reprezentacją dla innego komponentu – czyli

właśnie konsumenta. Inaczej mówiąc, w proponowanej tu koncepcji konsument zapewnia czemuś *rolę* bycia reprezentacją[21].

Druga różnica zachodząca między kategorią konsumenta u Millikan a pojmowaniem tej kategorii w rozwijanej tu koncepcji ma źródło w tym, jak w obydwu teoriach rozumie się naturę *funkcji*. Uznanie czegoś za konsumenta reprezentacji to pewna kwalifikacja funkcjonalna: być konsumentem to odgrywać określoną rolę. Dla Millikan (1984, 2002) pełnienie funkcji (tak zwanej funkcji właściwej) jest konstytuowane zależne od historii selekcyjnej tego, czemu taką funkcję przypisujemy. Z takiej perspektywy pewien komponent mechanizmu będzie pełnił funkcję konsumenta reprezentacji, jeśli rozpowszechnił się on jako cecha gatunkowa na drodze doboru naturalnego, to znaczy dlatego, iż to właśnie „konsumując" reprezentacje, zwiększał on poziom dostosowania – mierzony w „walucie" sukcesu reprodukcyjnego – poszczególnych jednostek tego gatunku. Tymczasem teoria MKM czerpie rozumienie funkcji z koncepcji wyjaśniania mechanistycznego (por. omówienie zagadnienia funkcji w mechanizmach zawarte w sekcji 2.1.1). Przypisując komponentom mechanizmu określone funkcje (operacje), mam więc na myśli *funkcje w sensie Cumminsa* (por. Cummins 1975). Funkcja w sensie Cumminsa jest określana przez wkład danego komponentu mechanizmu w umożliwianie zjawiska wyższego poziomu (zjawiska, za które odpowiada ten mechanizm). Konsument reprezentacji zawdzięcza zatem swój status funkcjonalny nie historii selekcyjnej, lecz temu, że właśnie będąc konsumentem, przyczynia się on do określonego zjawiska z wyższego poziomu organizacji[22].

---

[21] Nie zaprzeczam, iż konsument odgrywa także rolę przy determinowaniu treści reprezentacji. Nie jest to jednak zasadnicza czy pierwotna rola, jaką chcę przypisywać konsumentowi w prowadzonych tu rozważaniach. Mimo to w sekcji 4.2.2 poruszę także właśnie kwestię znaczenia konsumenta dla determinacji treści reprezentacji.

[22] Wydaje się, że takie podejście do natury funkcji ma pewne przewagi nad podejściem ewolucyjnym (prezentowanym przez Millikan). Załóżmy na potrzeby argumentacji, że niektóre powstałe w wyniku ewolucji przez dobór naturalny systemy poznawcze rzeczywiście wykorzystują mechanizmy wyposażone w konsumowane modele. Przyjmijmy na przykład, że w MKM są wyposażone ośrodkowe układy nerwowe lwów. Wyobraźmy sobie teraz, że gdzieś na ziemi

Przejdźmy teraz do drugiego warunku nakładanego na reprezentacyjną triadę:

(W2) Pełnienie przez nośnik reprezentacji roli wymienionej w (W1) jest w systematyczny sposób zależne od relacji zachodzącej między nośnikiem a przedmiotem reprezentacji. To znaczy:

---

w bagno uderza piorun i niezwykłym zbiegiem okoliczności powstaje w ten sposób idealny molekularny duplikat lwa: bagno-lew. Gdybyśmy rozumieli funkcje pełnione przez komponenty MKM jako funkcje właściwe w sensie Millikan, pojawiłyby się dwa problemy. Po pierwsze, ze względu na brak historii selekcyjnej, bagno-lew okazałby się nie być system reprezentacyjnym: nic w jego wnętrzu nie pełniłoby roli konsumenta (jak również nośnika) reprezentacji. Okazałoby się, w sposób wysoce kontrintuicyjny, że choć systemy poznawcze lwa i bagno-lwa działają *prima facie* identycznie, to jeden z nich korzysta z reprezentacji, a drugi – nie. Jest to konsekwencja, której zapewne chcielibyśmy uniknąć: wydaje się, że oba te systemy wykorzystują wewnętrzne, konsumowane modele w swoim działaniu. Po drugie, ten eksperyment myślowy wydaje się pokazywać, że funkcje u Millikan są przyczynowo epifenomenalne (Christensen, Bickhard 2002; Bickhard 2004b). W ujęciu Millikan o pełnieniu funkcji nie decydują własności przyczynowe posiadane przez komponenty mechanizmu w danym momencie, lecz jedynie historia tych komponentów. Fakt ten stoi w sprzeczności z wiarygodnym postulatem, że funkcje powinny być przyczynowo efektywne (Christensen, Bickhard 2002; Bickhard 2004b). Kiedy rozumiemy funkcje jako funkcje w sensie Cumminsa, żaden z dwóch wymienionych problemów w ogóle się nie pojawia. Będąc molekularnymi „bliźniakami", lew i bagno-lew charakteryzują się taką samą wewnętrzną, mechanistyczną architekturą; mechanistyczna (strukturalno-funkcjonalna) dekompozycja ich systemów poznawczych przyniesie te same rezultaty. Jedynym, co różni te organizmy, jest ich historia. W związku z tym nie ma żadnych przeciwwskazań by uznać, że także bagno-lew jest „wyposażony" w MKM. Ponadto funkcje w sensie Cumminsa wydają się przyczynowo efektywne. Ich atrybucja dokonuje się na podstawie sposobu, w jaki dany komponent (któremu przypisujemy funkcję) przyczynia się do zjawiska z wyższego poziomu. Tym samym komponenty realizują funkcje (operacje) w wyniku posiadanych własności przyczynowych (czyli na podstawie tego, jak interreagują z innymi komponentami mechanizmu). Lew i bagno-lew dysponują zatem mechanizmami wyposażonymi w komponenty pełniące funkcje konsumentów (i nośników) reprezentacji dlatego, że komponenty te posiadają aktualnie określone własności przyczynowe. Pełnienie funkcji konsumenta (nośnika) reprezentacji nie jest z takiej perspektywy przyczynowo epifenomenalne.

> (W2-a) Istnieje pewnego rodzaju relacja, która może zachodzić (lub nie) między nośnikiem a przedmiotem reprezentacji. (W2-b) Zachodzenie relacji wymienionej w (W2-a) to warunek konieczny tego, by nośnik poprawnie pełnił funkcję wymienioną w (W1). Jeśli relacja taka nie zachodzi, nośnik reprezentacji nie pełni funkcji wymienionej w (i) w sposób poprawny.

Warunek ten dotyczy relacji między nośnikiem a przedmiotem reprezentacji. Bardziej precyzyjnie, relacji tej dotyczy (W2-a), natomiast (W2-b) wiąże jej zachodzenie (lub nie) z określonym wpływem nośnika na konsumenta. Sądzę, że reprezentacje pojmowane jako konsumowane modele mentalne spełniają oba człony (W2). W zgodzie z (W2-a) proponowana tu koncepcja przypisuje zasadnicze znaczenie relacji zachodzącej między nośnikiem a przedmiotem reprezentacji. Relację tę w MKM stanowi podobieństwo strukturalne. Na przykład w samochodzie Cumminsa kluczowe znaczenie ma podobieństwo przestrzenno-strukturalne zachodzące między rowkiem wyżłobionym w tablicy (nośnikiem) a torem, po którym porusza się pojazd (przedmiotem reprezentacji). Jednocześnie w zgodzie z (W2-b) w MKM nośnik nie może pełnić swojej roli poprawnie, o ile nie będzie on strukturalnie podobny do przedmiotu reprezentacji. Kluczowe znaczenie ma tu konsument reprezentacji i to, jak jest on funkcjonalnie uzależniony od nośnika. Mechanizm korzystający z wewnętrznego modelu cechuje się zawsze organizacją sprawiającą, że konsument będzie poprawnie realizował swoje działanie tylko pod warunkiem, iż „otrzymuje" od nośnika wierną czy poprawną „wiadomość" na temat stanu, w jakim znajduje się (względnie: będzie się znajdował lub znajdowałby się w określonych kontrfaktycznych okolicznościach) przedmiot reprezentacji. Jeśli nośnik nie jest strukturalnie podobny do tego przedmiotu, nie będzie on poprawnie zastępował go dla konsumenta; zarazem jeśli relacja podobieństwa zachodzi, nośnik będzie wpływał na konsumenta, tak że ten ostatni będzie działać w sposób dostosowany do zachodzących okoliczności (do stanu, w jakim znajduje się przedmiot reprezentacji). Na przykład dopóki rowek wyryty w tablicy w samochodzie

Cumminsa zachowuje podobieństwo do toru, dopóty wpływa on na ster w taki sposób, że ten ostatni odpowiednio kieruje ruchem samochodu; kiedy jednak podobieństwo zostanie zaburzone, ster zacznie działać w sposób uniemożliwiający pojazdowi sprawne poruszanie się po torze.

Warto w tym miejscu przypomnieć pytanie postawione w sekcji 4.1.3 tego rozdziału: jak dokładne podobieństwo powinno zachodzić między nośnikiem a przedmiotem reprezentacji, abyśmy mieli do czynienia z poprawną reprezentacją? Jak bardzo wyidealizowany i niekompletny może być wewnętrzny model, aby nadal mógł być poprawny? Otóż dysponowanie pojęciem konsumenta pozwala na sformułowanie odpowiedzi na to pytanie. Tak jak niekompletna i do pewnego stopnia „przekłamana" mapa metra może być poprawna, dopóki korzystająca z niej osoba odnosi sukces nawigacyjny, tak niekompletny i do pewnego stopnia przekłamany model wewnętrzny może być poprawny, dopóki zapewnia konsumentowi poprawne działanie. Odpowiednio zaprojektowany (przez konstruktora bądź dobór naturalny) konsument będzie „tolerował" braki, uproszczenia i przeinaczenia modelu; to znaczy będzie on działał w sposób dostosowany do przedmiotu reprezentacji nawet wtedy, gdy jego działaniami przewodzi nośnik wykazujący jedynie ograniczone podobieństwo do tego przedmiotu[23]. Model (nośnik modelu) przestaje

[23] Można zilustrować tę myśl na przykładzie samochodu Cumminsa. Jak pamiętamy, pojazd ten ma za zadanie przemierzyć tor w kształcie litery „S". Aby to zrealizować, pojazd wykorzystuje wewnętrzną, S-kształtną mapę, która jest w odpowiedni sposób podłączona do steru. Możemy sobie jednak wyobrazić sytuację, w której układ nośnik–konsument (mapa–ster) dopuszcza pewien stopień „swobody" jeśli chodzi o podobieństwo między nośnikiem a przedmiotem reprezentacji (mapą a torem). Oznacza to, że nie tylko mapa idealnie czy maksymalnie odzwierciedlająca tor, ale także mapy w mniejszym stopniu strukturalnie podobne do toru „nadają" się do sterowania pojazdem. Wykorzystując takie mniej wierne mapy, pojazd będzie czasem niebezpiecznie zbliżał się do krawędzi, skręcał pod nie do końca odpowiednim kątem i w nie do końca odpowiednich miejscach, jednak nie doprowadzi to do (zbyt mocnych) kolizji z krawędziami i nie uniemożliwi pojazdowi pokonania toru. Każda z tych map będzie mogła zostać uznana za poprawną (poprawny model toru), ponieważ każda z nich zapewnia sukces konsumentowi (sterowi, który zawiaduje ruchem pojazdu).

być poprawny wtedy, gdy jego braki i przeinaczenia są na tyle duże, że funkcjonowanie konsumenta zostaje zaburzone (nie odnosi on sukcesu).

Powyższa konstatacja pokazuje pośrednio, że z bronionego tu punktu widzenia relacja między nośnikiem a przedmiotem reprezentacji jest tak samo niezbędna dla posiadania statusu mechanizmu reprezentacyjnego, jak relacja między nośnikiem a konsumentem reprezentacji. W rozdziale 3 (sekcja 3.2.1) wspomniałem, że niektórzy przedstawiciele ucieleśnionego/enaktywnego nurtu w naukach kognitywnych postulują, iż teorie reprezentacji w kognitywistyce powinny skupiać się na roli czy funkcji, jaką reprezentacje spełniają dla działających w środowisku, ucieleśnionych systemów poznawczych. Koncepcje tego rodzaju miałyby zastąpić bardziej klasyczne teorie, koncentrujące się – jak to ujmuje Mark Bickhard (2004a, 2004b, 2009) – na reprezentacji rozumianej jako „kodowanie" (*encoding*), to znaczy przede wszystkim jako „korespondencja" między reprezentacją (nośnikiem) a tym, co reprezentowane (przedmiotem). Jednym z przykładów podejścia alternatywnego w stosunku do tego opartego na kodowaniu jest omówiona w poprzednim rozdziale teoria reprezentacji jako przewodników działań (Anderson, Rosenberg 2008). Sam wspomniany Bickhard (2004a, 2004b, 2009) z kolei rozwija koncepcję reprezentacji skoncentrowaną wokół pojęcia *interakcji*. Mówiąc w zarysie, z punktu widzenia tego autora reprezentacje stanowią rozwiązanie określonego problemu związanego z kontrolowaniem interakcji systemu poznawczego ze środowiskiem. Otóż odpowiednio złożone rodzaje systemów poznawczych mogą potencjalnie zaangażować się w jednym momencie w wiele alternatywnych interakcji ze środowiskiem. Jednak tylko niektóre z tych ostatnich mogą odnieść sukces, biorąc pod uwagę zachodzące warunki zewnętrzne. Reprezentacje miałyby stanowić wskazówkę (*indication*) potencjalnego sukcesu różnych alternatywnych, możliwych do nawiązania w danych okolicznościach interakcji. Pozwalałoby to systemowi na wybór optymalnej (najbardziej pożądanej) w tych okolicznościach opcji. Zasadnicza rola reprezentacji polega zatem według Bickharda na umożliwianiu *preselekcji działań*.

Uważam, że traktowanie działaniowego (interakcyjnego) podejścia w teorii reprezentacji oraz podejścia opartego na kodowaniu (korespondencji) jako wykluczających się alternatyw stanowi złe postawienie sprawy (por. także Gładziejewski 2015). Pierwsze z tych podejść jest narażone na sprowadzenie reprezentacji do relacji między konsumentem a przedmiotem reprezentacji. Reprezentacje sprowadzone do użycia (interakcji, przewodzenia działaniami i tak dalej) okazują się jednak – jak to próbowałem wykazać w rozdziale 3 na przykładzie teorii Andersona i Rosenberga – tracić swój funkcjonalny status. Drugie podejście „redukuje" z kolei reprezentację do relacji między jej nośnikiem a przedmiotem. Reprezentacje sprowadzone do korespondencji między nośnikiem a przedmiotem wydają się jednak funkcjonalnie jałowe; nie wiadomo, w jaki sposób pełnią one funkcję reprezentacji w szerszym mechanizmie czy systemie. Twierdzę, że skrojona na potrzeby kognitywistyki koncepcja reprezentacji (wyjaśniania reprezentacyjnego) musi stanowić syntezę obu tych podejść. Dopiero umiejętne połączenie tych dwóch perspektyw może stanowić podstawę dla teorii reprezentacji, która czyni zadość wymogowi opisu zadań. Teoria taka musi powoływać się zarówno na relację nośnik–przedmiot, jak i na rolę, jaką ta relacja odgrywa w kontrolowaniu działań czy interakcji.

Broniona tu koncepcja mechanizmów reprezentacyjnych jako wyposażonych w wewnętrzne, konsumowane modele stanowi właśnie tego rodzaju syntezę. Wewnętrzne modele nie sprowadzają się ani do użycia (zależności konsument–przedmiot), ani do korespondencji (relacji nośnik–przedmiot). Z perspektywy teorii MKM nie wchodzi w grę mówienie o wewnętrznym modelu, który nie jest *wykorzystywany* jako model właśnie. Mówiąc inaczej, „modele", które nie są konsumowane, nie powinny być w ogóle klasyfikowane jako modele. W kontekście MKM reprezentacja jest niemożliwa, jeśli nie przewodzi działaniami pewnego mechanizmu (przez przewodzenie działaniami konsumenta/konsumentów). Komponent mechanizmu staje się nośnikiem reprezentacji nie tylko dlatego, że znajduje się on w relacji podobieństwa strukturalnego do przedmiotu reprezentacji, ale także dlatego, że jest on wykorzystywany jako reprezentacja przez konsumenta. Broniona tu teoria nie stoi zatem w sprzecz-

ności z fundamentalną obserwacją Ludwiga Wittgensteina (2000), że obrazy reprezentują tylko dzięki użyciu. Pamiętajmy jednak, że jednocześnie dla bycia *konsumentem* jest konstytutywne, iż stanowi on strukturę eksploatującą relację między nośnikiem a przedmiotem reprezentacji. Konsument, który nie jest odpowiednio powiązany z nośnikiem, traci swój status; staje się „zwykłym" komponentem mechanizmu, którego aktywność nie opiera się na wykorzystywaniu reprezentacji. W świetle koncepcji MKM reprezentacje są więc przewodnikami działań, jednak przewodnikami specyficznymi – takimi, które wykorzystują relację podobieństwa zachodzącą między modelem a tym, co przez ten model reprezentowane[24]. W ten sposób proponowana tu koncepcja łączy podejście interakcyjne/działaniowe – z odwołującym się do reprezentacji jako kodowania.

4.2.2. Teoria mechanizmów wykorzystujących konsumowane modele. Zarzuty i odpowiedzi

W poprzedniej sekcji sformułowałem zasadnicze twierdzenia proponowanej tu koncepcji mechanizmów reprezentacyjnych. Nie ulega jednak wątpliwości, że teoria ta w przedstawionej wyżej postaci natrafia na określone problemy. Niektóre jej tezy mogą rodzić wątpliwości i pytania, a inne mogą wydawać się podatne na zarzuty. W tej sekcji chcę wymienić te problemy – a w każdym razie te spośród nich, które uważam za najpoważniejsze i pilnie wymagające rozwiązania –

[24] Zgadzam się z Bickhardem (2004a, 2004b, 2009), że reprezentacje są w systemie po to, by regulować jego interakcje ze środowiskiem. Umożliwianie antycypacji skutków alternatywnych działań stanowi jeden (szczególnie podkreślany przez samego Bickharda) ze sposobów, w jakie reprezentacje mogą regulować te interakcje. Jednak z bronionej przeze mnie perspektywy – reprezentacja umożliwia antycypację skutków alternatywnych działań wtedy, gdy stanowi uruchomiony *off-line* model potencjalnych interakcji organizm–środowisko. Bez dodania tezy o takiej „modelowej" czy symulacyjnej podstawie interakcji, pozostajemy tylko ze złożonymi interakcjami ze środowiskiem, takimi które *prima facie* można wyjaśnić bez powoływania się na reprezentacje. Inaczej mówiąc, do bycia reprezentacją nie wystarcza samo umożliwianie antycypacji czy predykcji. Reprezentacje umożliwiają predykcje dzięki temu, że są modelami tego, co tej predykcji podlega (por. Gładziejewski 2015).

oraz zaproponować odpowiedzi. Moim celem nie będzie jednak tylko obrona koncepcji mechanizmów reprezentacyjnych jako mechanizmów wyposażonych w wewnętrzne, konsumowane modele. Udzielenie odpowiedzi na wątpliwości i zarzuty ma jednocześnie stanowić okazję do uzupełnienia, rozjaśnienia i doprecyzowania bronionej tu teorii. Poniżej znajduje się lista przedstawiająca najsilniejsze i najbardziej znaczące (przynajmniej *prima facie*) problemy stojące przed teorią MKM, a także proponowane ich rozwiązania.

a) Teoria MKM jest czysto funkcjonalna, nie podaje szczegółów strukturalnych

Natura problemu: Wyjaśnienie jakiegoś zjawiska przez wskazanie odpowiedzialnego za nie mechanizmu wymaga wykonania zarówno funkcjonalnej, jak i strukturalnej dekompozycji danego systemu. Nie możemy zatem powiedzieć, że dysponujemy takim wyjaśnieniem, o ile nie potrafimy przypisać wyróżnionych działań (wynik dekompozycji funkcjonalnej) poszczególnym fizycznym komponentom mechanizmu (wynik dekompozycji strukturalnej). Tymczasem proponowana tu koncepcja mechanizmów reprezentacyjnych jako MKM skupia się jedynie na organizacji *funkcjonalnej* owych mechanizmów. Nie mówi ona nic o komponentach (na przykład o ich lokalizacji w ośrodkowym układzie nerwowym), które miałyby rzeczywiście realizować funkcję nośnika czy konsumenta reprezentacji. Jak mogę mówić zatem o teorii *mechanizmów* reprezentacyjnych, jeśli odwołuję się do dekompozycji funkcjonalnej, której nie towarzyszy dekompozycja strukturalna?

Odpowiedź: Aby odpowiedzieć na powyższy zarzut, warto raz jeszcze podkreślić, na czym polega cel teoretyczny prowadzonych tu rozważań. Nie jestem zainteresowany dostarczeniem mechanistycznego, reprezentacyjnego wyjaśnienia jakiegoś określonego zjawiska poznawczego. Gdyby to było moim celem, brak dekompozycji strukturalnej w istocie stanowiłby poważny problem, jednak nie temu ma służyć przedstawiana tu teoria. Koncepcja MKM jest rozwijana w kontekście *metaprzedmiotowego* projektu zmierzającego do sformułowania ogólnej koncepcji mechanistycznego wyjaśnia-

nia reprezentacyjnego. Chcę zatem podać możliwie ogólne kryteria bycia mechanizmem reprezentacyjnym, kryteria, które mogą zostać spełnione (na poziomie przedmiotowym) przez wiele różnych mechanizmów poznawczych odkrywanych w ramach nauk kognitywnych. Siłą rzeczy koncepcja MKM musi być zatem sformułowana w sposób stosunkowo abstrakcyjny. Właśnie z tego powodu zamiast kompletnego opisu jakiegoś konkretnego mechanizmu (jego komponentów, ich działań i wewnętrznej organizacji), proponuję *szkic* mechanizmu. Jest to zbieżne z omówioną w rozdziale 2 (sekcja 2.3.2) ideą, iż modele funkcjonalne stanowią wstępną, szkicową formę pełnoprawnych wyjaśnień mechanistycznych. Uważam, że przedstawienie takiego funkcjonalnego szkicu mechanizmu reprezentacyjnego dobrze wpisuje się w moje cele teoretyczne. Proponuję abstrakcyjny, funkcjonalny zarys mechanistycznego wyjaśnienia reprezentacyjnego, który może docelowo zostać uzupełniony szczegółami strukturalnymi.

b) Podobieństwo nie może stanowić podstawy dla reprezentacji
Natura problemu: Broniona tu koncepcja mechanizmów wykorzystujących modele może zostać potraktowana jako szczególny wariant pewnej ogólnej filozoficznej tezy, iż reprezentowanie opiera się na relacji podobieństwa. Jednak wszelkie teorie czyniące podobieństwo podstawą dla reprezentacji są podatne na dwa dyskwalifikujące zarzuty:

> **1. Zarzut z logicznych własności podobieństwa:** Relacja podobieństwa obiektu *x* do obiektu *y* ma inne własności logiczne, niż relacja reprezentowania obiektu *x* przez obiekt *y* (Goodman 1976; Suárez 2003; Miłkowski 2013: 150–151). Konkretnie: (1) podobieństwo jest samozwrotne, a reprezentowanie – nie; (2) podobieństwo jest symetryczne, a reprezentowanie – nie. Reprezentowanie *x* przez *y* nie może być zatem tym samym, co podobieństwo *x* do *y*[25].

[25] Warto od razu zauważyć, że zarzut odwołujący się do symetryczności relacji podobieństwa uderza w koncepcję reprezentacji opartą na podobieństwie

2. Zarzut z wszechobecności podobieństwa: Wszystko we Wszechświecie jest w jakiś sposób czy pod jakimś względem podobne do czegoś innego. Na przykład dowolny obiekt fizyczny jest podobny do dowolnego innego obiektu fizycznego, bo podziela z nim własność polegającą na posiadaniu (jakiegoś) położenia czasoprzestrzennego. To samo dotyczy podobieństwa strukturalnego. Jeśli tylko odpowiednio wyróżnimy elementy i relacje w dwóch systemach czy procesach, może okazać się, że telewizor jest strukturalnie podobny do ciała ludzkiego (Bartels 2006), albo że przebieg partii gry w szachy jest strukturalnie podobny do przebiegu wojny sześciodniowej (Fodor 1981: 207). Wydaje się jednak, że w żadnym z wymienionych przypadków nie mamy do czynienia z reprezentowaniem. Przebieg wojny sześciodniowej nie reprezentuje przebiegu żadnej partii szachów (chyba że ktoś go tak idiosynkratycznie wykorzysta). Co więcej, jeśli reprezentowanie opiera się na podobieństwie strukturalnym, to komputerowa symulacja (dynamiczna S-reprezentacja) przebiegu partii szachów może okazać się zarazem symulacją przebiegu wspomnianej wojny. To kolejna niewygodna konsekwencja teorii, która podobieństwo uznaje za podstawę reprezentacji. Zwolennik takiego podejścia staje przed jedną z dwóch możliwości. Po pierwsze, może on wycofać się ze swojej teorii (por.: Goodman 1976; Calvo, García 2009; Wittgenstein 2000). Po drugie, może on zaakceptować wysoce kontrintuicyjne konsekwencje swojej teorii i przyjąć, że każde podobieństwo jest przypadkiem reprezentowania. To jednak będzie implikować panreprezentacjonizm, a wraz z nim – trywializację eksplanacyjną pojęcia reprezentacji. Teoria reprezentacji odwołująca się do podobieństwa okazuje się zatem albo fałszywa, albo poznawczo bezwartościowa.

strukturalnym tylko wtedy, gdy owo podobieństwo przyjmuje swoją najmocniejszą postać, czyli izomorfizm (Bartels 2006). Gdy przyjmuje ono postać homomorfizmu lub izomorfizmu osadzonego (por. sekcja 4.1.3), podobieństwo strukturalne nie jest relacją symetryczną.

Odpowiedź: Odróżnijmy dwie możliwe interpretacje tezy, że reprezentowanie jest „oparte na podobieństwie". Zgodnie z silną interpretacją podobieństwo *wystarcza* dla reprezentacji, to znaczy zachodzenie podobieństwa między *x* a *y* stanowi warunek wystarczający do tego, aby *x* reprezentowało *y*. Zgodnie ze słabą interpretacją relacja podobieństwa jest w jakiś sposób *ważna* (na przykład eksplanacyjnie) dla reprezentacji, jednak nie stanowi warunku wystarczającego. Silna interpretacja okazuje się podatna na opisane wyżej zarzuty, jednak słaba – niekoniecznie. Otóż teoria wewnętrznych reprezentacji jako modeli mentalnych stanowi słabą interpretację tezy o podobieństwie jako podstawie reprezentacji. Zobaczmy, dlaczego tak jest i w jaki sposób fakt ten pozwala bronionej tu koncepcji uniknąć obu wymienionych wyżej zarzutów.

Koncepcja MKM nie zawiera ani nie implikuje twierdzenia, że podobieństwo strukturalne stanowi warunek wystarczający reprezentacji. Relacja podobieństwa, a konkretnie podobieństwa strukturalnego, jest bez wątpienia bardzo ważna dla bycia modelem mentalnym. Jednak bycie modelem nie sprowadza się do podobieństwa. Modele w moim ujęciu są w nieunikniony sposób strukturami używanymi właśnie jako modele przez konsumentów reprezentacji. Bycie „konsumowanym" przez jakiś komponent mechanizmu pozostaje dla modelu tak samo istotne czy konstytutywne, jak relacja podobieństwa strukturalnego. Podobieństwa strukturalne stają się podobieństwami *reprezentującymi* tylko wtedy, kiedy są one stosownie wykorzystywane przez konsumentów.

Odnieśmy się zatem do dwóch wymienionych wyżej zarzutów:

Ad 1. Zarzut z logicznych własności podobieństwa: Zgodnie z bronioną tu koncepcją, bycie reprezentacją opiera się nie tyle na podobieństwie, co raczej na byciu (konsumowanym) modelem. W odróżnieniu od podobieństwa – bycie przez *x* modelem *y* (1) nie jest relacją samozwrotną, jak również (2) nie jest relacją symetryczną. Relacja bycia modelem nie jest samozwrotna, ponieważ modele nie zastępują dla konsumentów samych siebie, lecz to, do czego konsumenci dostosowują (właśnie dzięki wykorzystaniu modeli) własne funkcjonowanie. Nie można tu mówić

zarazem o relacji symetrycznej: to, co modelowane przez pewien model A – przedmiot reprezentacji – samo nie staje się (*ceteris paribus*[26]) modelem B, którego przedmiotem reprezentacji byłby (symetrycznie) właśnie model A. Tym samym relacja bycia modelem nie ma własności logicznych, których nie mogłaby podzielać z relacją reprezentowania.

Ad 2. Zarzut z wszechobecności podobieństwa: Zgodnie z koncepcją MKM tylko bardzo niewielki podzbiór zachodzących we Wszechświecie podobieństw strukturalnych to podobieństwa reprezentujące. Reprezentują bowiem tylko te podobieństwa, które są odpowiednio wykorzystywane przez konsumentów reprezentacji[27]. Nie powstaje zatem beznadziejna alternatywa, przed którą miałby rzekomo zwolennika teorii MKM stawiać zarzut z wszechobecności podobieństwa. Po pierwsze, nie trzeba się wycofywać z koncepcji modeli wewnętrznych, ponieważ nie opiera się ona na twierdzeniu, jakoby podobieństwo miało być wystarczającym warunkiem reprezentacji. Po drugie, na gruncie

[26] Dodanie tu zastrzeżenia c*eteris paribus* jest motywowane następująco. Możemy wyobrazić sobie sytuację, w której pewien mechanizm, nazwijmy go ME1, zawiera nośnik reprezentacji (model mentalny) N1, który to jest konsumowany przez komponent K1, a którego przedmiot reprezentacji stanowi jakaś zewnętrzna struktura P. Przyjmijmy jednak, że okazuje się, iż P sam stanowi komponent innego mechanizmu, ME2. Podobieństwo strukturalne zachodzące między P a N1 jest konsumowane wewnątrz ME2 przez jakiś inny komponent ME2, K2. P przewodzi w ME2 działaniami K2 względem N1, tak jak N1 przewodzi w ME1 działaniami K1 względem P. Innymi słowy, okazuje się, że N1 i P stanowią swoje „symetryczne" modele (N1 jest modelem P, a P – modelem N1). Broniona tu teoria jak najbardziej dopuszcza takie sytuacje jako możliwe. Czy rehabilituje to jednak argument odwołujący się do symetrii jako własności logicznej? Nie. Opisany przypadek stanowi pewien *możliwy nomologicznie/empirycznie* stan rzeczy. Niekiedy może *a posteriori* okazać się po prostu, że mamy do czynienia z dwiema strukturami, które stanowią swoje „symetryczne" modele. Nomologiczna czy empiryczna możliwość zachodzenia takich sytuacji nie czyni jednak symetrii *logiczną* własnością relacji bycia modelem.

[27] Analogiczną procedurę można zastosować rzecz jasna do reprezentacji zewnętrznych. W tym przypadku reprezentują tylko te podobieństwa, które są w odpowiedni sposób wplecione w praktyki interpretacyjne podmiotów intencjonalnych, to znaczy – są interpretowane jako reprezentacje.

koncepcji MKM nie stoimy przed zagrożeniem panreprezentacjonizmu. Kiedy mówimy, że coś stanowi model mentalny, wyrażamy nietrywialną tezę o funkcjonowaniu tego czegoś w ramach szerszego mechanizmu.

c) Koncepcja MKM nie daje odpowiedzi na pytanie o sposób determinacji treści intencjonalnej

Natura problemu: Reprezentacje z istoty czegoś dotyczą, są o czymś. Każda z nich posiada treść, czyli warunki poprawności – które powinny być spełnione, jeśli reprezentacja ta ma zostać zakwalifikowana jako poprawna lub prawdziwa. Wydaje się, że ta zasadnicza idea powinna mieć także zastosowanie w przypadku reprezentacji postulowanych przez kognitywistów. Także one powinny więc – aby w ogóle zasługiwać na miano reprezentacji – posiadać treść. Tymczasem koncepcja mechanizmów korzystających z wewnętrznych modeli milczy na temat treści. Koncentruje się ona na funkcjach modeli mentalnych w mechanizmach, jednak nie daje odpowiedzi na pytanie o to, co stanowi czynnik determinujący treść intencjonalną tych modeli. Teoria MKM może być zatem uznana za fundamentalnie niekompletną.

Odpowiedź: W rzeczy samej teoria MKM nie zawiera bezpośredniej odpowiedzi na pytanie o sposób determinacji treści intencjonalnej. Pamiętajmy jednak, że jest to jak najbardziej zgodne z przeznaczeniem tej teorii. Koncepcja mechanizmów reprezentacyjnych ma w zamierzeniu stanowić teorię funkcjonowania jako reprezentacja w ramach systemu poznawczego, a nie teorię treści. Mimo to chcę pokazać, że ta koncepcja dysponuje zasobami konceptualnymi, dzięki którym może się potencjalnie „uporać" także z zagadnieniem treści intencjonalnej[28].

Aby zobaczyć, w jaki sposób na podstawie teorii MKM można sformułować koncepcję treści, należy odróżnić dwa możliwe spo-

[28] Fakt ten jest szczególnie istotny, dlatego że, jak stwierdziłem w rozdziale 3 (sekcja 3.1.1), problem funkcji reprezentacji i problem treści nie są całkowicie odrębne i autonomiczne.

soby pojmowania *przedmiotu reprezentacji*. Z jednej strony może on być rozumiany jako to, do czego jest *aplikowany* model mentalny (nośnik reprezentacji) w aktualnej sytuacji. Nazwijmy przedmiot w tym znaczeniu „przedmiotem aplikacji" (por. kategoria *target* w: Cummins 1996: 113–122; por. także: Ramsey 2007: 93–96; Herschbach 2012). Z drugiej strony – może być rozumiany jako *warunki poprawności* (spełnienia) modelu, to znaczy warunki, które powinny zajść, aby model (nośnik reprezentacji) w danej sytuacji poprawnie spełnił swoją rolę funkcjonalną. Przedmiot reprezentacji w takim znaczeniu nazwijmy „docelowym".

Powyższe odróżnienie można zilustrować, odwołując się raz jeszcze do bardzo prostego, poglądowego przykładu, jaki stanowi samochód Cumminsa. Wyobraźmy sobie, że porusza się on, wykorzystując wewnętrzną mapę (tablicę), w której wydrążono kształt litery „S". Załóżmy jednocześnie, że tor, po którym porusza się pojazd, ma kształt *odwróconej* litery „S". Samochód startuje, szybko uderza w krawędź toru, wykonuje serię chaotycznych ruchów i w ostateczności nie udaje mu się przemierzyć toru. Jest to przykład sytuacji, w której dochodzi do rozbieżności między *przedmiotem aplikacji* a *przedmiotem docelowym*. Samochód aplikuje wewnętrzną mapę do toru o kształcie odwróconej litery „S". Jest to bowiem tor, w którym rzeczywiście czy aktualnie znalazł się pojazd. Innymi słowy, znajdujący się w samochodzie ster *de facto* wykorzystuje wewnętrzną mapę do nawigowania ruchem samochodu właśnie po takim torze. Tor w kształcie odwróconego „S" stanowi więc tutaj przedmiot aplikacji. Jednak przedmiotem docelowym jest w naszym przypadku tor o kształcie litery „S". Dlaczego? Jeśli chcemy teraz wyznaczyć przedmiot docelowy, powinniśmy wziąć pod uwagę kształt wewnętrznej mapy (rowka wydrążonego w tablicy), a następnie zadać pytanie: po jakim torze samochód powinien się poruszać, aby ta wewnętrzna mapa funkcjonowała poprawnie, czyli zgodnie ze swoim przeznaczeniem? Biorąc pod uwagę to, w jaki sposób są ze sobą powiązane (przyczynowo oraz funkcjonalnie) nośnik i konsument reprezentacji, wiemy, że jeśli nośnik (rowek w tablicy) przyjmuje kształt litery „S", to konsument (ster) będzie generował ruch samochodu układający się właśnie w kształt litery „S". Aby samochód mógł skutecz-

nie przemierzyć tor, także tor powinien układać się (w przybliżeniu) w kształt „S". Właśnie z tego powodu tor w kształcie „S" stanowi w omawianym przykładzie przedmiot docelowy. Powtórzę więc: mamy tu do czynienia z sytuacją, w której przedmiot aplikacji reprezentacji jest inny niż jej przedmiot docelowy. Ujmując rzecz inaczej, można powiedzieć, że model, jakim posługuje się nasz samochód, jest *niepoprawny* lub *błędny*.

Na podstawie powyższych rozstrzygnięć *treść* modelu mentalnego proponuję utożsamić z jego przedmiotem docelowym. Przedmiot docelowy to bowiem nie tyle jakiś realny, fizyczny obiekt w środowisku, co raczej *warunki poprawności* modelu. To coś, do czego powinien być podobny model, aby był on poprawny. Nietrudno teraz rozstrzygnąć kwestię, jak jest determinowana treść modeli wewnętrznych. Pytanie o sposób determinacji treści modeli można bowiem przełożyć na pytanie o sposób determinacji ich przedmiotów docelowych. W jaki sposób jest wyznaczany przedmiot docelowy modelu? Otóż określany jest on przez warunki poprawnego funkcjonowania układu nośnik–konsument. Nośnik poprawnie spełnia swoją funkcję w mechanizmie tylko wtedy, gdy wykazuje podobieństwo (w wymaganym zakresie) do przedmiotu docelowego. Jednak poprawne działanie nośnika jest pochodne względem jego wpływu na konsumenta reprezentacji. Konsument zaś funkcjonuje poprawnie tylko wtedy, gdy jego aktywność zostaje dostosowana do przedmiotu docelowego. Przedmiotem docelowym (treścią) jest zatem przedmiot (proces, stan rzeczy i tak dalej), do którego *powinien być podobny nośnik*, gdyby miał on (nośnik) poprawnie spełniać swoją funkcję; jednocześnie jest to przedmiot, do którego byłoby w takiej sytuacji dostosowane działanie *konsumenta*[29]. Treść reprezenta-

[29] Tezę tę można zilustrować jeszcze w następujący sposób. Wewnętrzna S-kształtna mapa, w którą wyposażono samochód Cumminsa, może wykazywać przestrzenne podobieństwo do wielu obiektów. Może ona być podobna kształtem do węża wspinającego się po drzewie w afrykańskiej dżungli, do linii narysowanej przez pewne dziecko na lekcji plastyki, do fragmentu polskiej linii brzegowej nad Bałtykiem, do S-kształtnego toru i tak dalej. Co decyduje o tym, że mapa reprezentuje tor (jako przedmiot docelowy), a nie węża, rysunek dziecka, fragment linii brzegowej czy dowolny inny obiekt, do którego jest ona podobna?

cji jest więc wyznaczana przez warunki sukcesu konsumenta, które to są zarazem takie same, jak warunki sukcesu nośnika. Z takiej perspektywy modele niepoprawne czy błędne to modele aplikowane w sytuacji, w której nie zostają zrealizowane warunki poprawnego działania układu nośnik–konsument. Inaczej mówiąc, model niepoprawny lub błędny (błędna reprezentacja) to model aplikowany do przedmiotu, który jest inny niż przedmiot docelowy.

Przedstawiona wyżej propozycja stanowi specyficzną odmianę teleosemantyki, a dokładniej, tak zwanej semantyki konsumentów (*consumer semantics*; por. Millikan 1986; 2002)[30]. Sądzę, że opisana tu koncepcja determinacji treści jest *prima facie* wiarygodna oraz odporna na co bardziej znaczące zarzuty, jakie można by wobec niej wysunąć[31]. Cel mojej pracy to jednak nie stworzenie komplekso-

Czynnikiem decydującym okazuje się tu oczywiście *funkcja* mapy: służy ona do nawigowania ruchem pojazdu po torze. Atrybucja tego rodzaju funkcji opiera się zaś na tym, że to właśnie od podobieństwa nośnika (mapy) do toru (a nie węża, dziecięcego rysunku czy fragmentu polskiej linii brzegowej) zależy sukces konsumenta reprezentacji (steru).

[30] Moja propozycja różni się w stosunku do szeregu klasycznych teleologicznych teorii treści przede wszystkim tym, że inaczej pojmuje się w niej naturę funkcji (to znaczy jako funkcje w sensie Cumminsa, a nie jako funkcje właściwe w sensie Millikan; por. przypis 22).

[31] Na przykład wspomniany Cummins stwierdziłby być może (por. 1996: 29–52), iż (1) naszkicowana tu teoria głosi, że treść reprezentacji jest określana przez użycie, a przecież (2) na gruncie takiej semantyki nie są możliwe reprezentacje fałszywe. Reprezentacje nie mogą bowiem być fałszywe, jeśli użycie determinuje ich treść. Wszakże na gruncie takiej teorii reprezentacja zawsze będzie reprezentować to, do czego jest aplikowana (to znaczy do reprezentowania czego jest używana). Nie mamy zatem narzędzi potrzebnych do tego, aby odróżnić przedmiot aplikacji od przedmiotu docelowego: „docelowość" jest konstytuowana przez aplikację.

Uważam, że powyższy zarzut nie jest uprawniony. Teoria mechanizmów reprezentacyjnych jako wyposażonych w konsumowane modele pozwala odróżnić przedmiot aplikacji od przedmiotu docelowego. Dostarcza ona narzędzia konceptualne pozwalające na pokazanie, iż istnieje różnica między przedmiotem, do którego model jest *de facto* aplikowany, od przedmiotu, do którego (moglibyśmy powiedzieć: *de iure*) powinien on być aplikowany, biorąc pod uwagę funkcjonalną strukturę mechanizmu. Ten pierwszy przedmiot będzie zatem przedmiotem aplikacji, a drugi – przedmiotem docelowym. Jeżeli dochodzi do rozbieżności między nimi, mamy do czynienia z reprezentacją błędną.

wej teorii treści modeli mentalnych. Chciałem tu jedynie pokazać, że stworzenie takiej teorii okazuje się na gruncie koncepcji MKM wykonalne. Uważam, że przedstawiony powyżej zarys pokazuje, iż tak rzeczywiście jest. Na powyższych ustaleniach chcę jednak zakończyć omawianie tego zagadnienia.

d) Status eksplanacyjny modeli mentalnych nie różni się od statusu eksplanacyjnego receptorów

Natura problemu: Jeśli prowadzone tu rozważania są poprawne, to istnieje zasadnicza różnica między statusem eksplanacyjnym modeli mentalnych a statusem eksplanacyjnym reprezentacji pojętych jako receptory. Te pierwsze spełniają wymóg opisu zadań, a drugie – nie. Ktoś może jednak wyrazić następującą wątpliwość (por. Ramsey 2007: 189–203). Na pierwszy rzut oka jedynym, co może odróżniać modele mentalne od receptorów, pozostaje relacja między nośnikiem a przedmiotem reprezentacji. Zauważmy jednakże, że to samo można by powiedzieć o odpowiednich reprezentacjach pozamentalnych. Wydaje się, że jedyna różnica między zewnętrznymi indeksami a ikonami dotyczy tego, jak nośnik jest powiązany z przedmiotem reprezentacji. W przypadku reprezentacji indeksowych mamy do czynienia ze współzmiennością (czy zależnością przyczynową), natomiast w przypadku ikon – z podobieństwem. Jednak zarówno pozamentalne ikony, jak i indeksy w niekontrowersyjny sposób mają funkcjonalny status reprezentacji. Dlaczego w przypadku reprezentacji mentalnych nie może być podobnie? Dlaczego przypisuję status reprezentacji konsumowanym modelom, nie czyniąc tego samego w przypadku mechanizmów, które wykorzystują coś, co moglibyśmy nazwać „konsumowanymi receptorami"? Czy magnetosom nie może zostać uznany za nośnik reprezentacji, którego konsumentem jest układ ruchowy bakterii (por. sekcja 3.3.1)? Na jakiej podstawie twierdzę, że w przypadku reprezentacji mentalnych zmiana *relacji* nośnik–przedmiot (z podobieństwa strukturalnego na współzmienność) skutkuje utratą *funkcjonalnego* statusu reprezentacji? Pytania te są szczególnie ważkie, jeśli weźmiemy pod uwagę, że posługiwanie się receptorowym pojęciem reprezentacji jest tak powszechne w kognitywistyce. Skoro odmawiam takim strukturom miana repre-

zentacji, powinienem mieć ku temu solidne podstawy. Tymczasem może wydawać się, że narzędzia konceptualne służące do sformułowania teorii mechanizmów korzystających z wewnętrznych modeli mogłyby równie dobrze, i to wcale nie *ad hoc*, posłużyć do sformułowania koncepcji mechanizmów korzystających z reprezentacji opartych na współzmienności. Czy nie potraktowałem wewnętrznych receptorów w sposób nieuczciwy?

Odpowiedź: Sądzę, że funkcjonalne różnice między wewnętrznymi modelami a receptorami są rzeczywiście na tyle znaczące, aby zagwarantować tym pierwszym, ale już nie drugim status reprezentacji. Żeby to wykazać, powołam się raz jeszcze na rozważania Ramseya (2007: 194–199). Autor ten bezpośrednio skonfrontował reprezentacje jako modele z (rzekomymi) reprezentacjami receptorowymi. Wyobraźmy sobie zatem za Ramseyem trzy różne samochody. Każdy z nich pokonuje S-kształtny tor, nie mając w środku kierowcy. Każdy z tych pojazdów działa jednak na nieco innej zasadzie:

> **Samochód A** ma z przodu dwie wypustki; jedna jest skierowana w lewą, a druga w prawą stronę. Wypustki są połączone z kierownicą samochodu. Kiedy lewa wypustka styka się z lewą krawędzią toru, przekręca ona kierownicę w przeciwną stronę, sprawiając, że samochód skręca w prawo. Na podobnej zasadzie wypustka położona z prawej strony zwraca samochód w lewo, kiedy zbliża się on do prawej krawędzi toru. W ten właśnie sposób samochód A pokonuje tor.
> **Samochód B** to samochód Cumminsa. Pokonuje on tor, wykorzystując wewnętrzny, konsumowany model toru.
> **Samochód C** przemierza tor, odbijając się od krawędzi do krawędzi. Kiedy jego koła stykają się z krawędzią toru, odwracają się one w drugą stronę. Pojazd jedzie wtedy w przeciwnym kierunku, dopóki znowu nie uderzy w krawędź, by w efekcie zwrócić się w drugą stronę i tak dalej. Ostatecznie pokonuje on jednak tor.

Samochód C jest w niekontrowersyjny sposób pozbawiony wewnętrznych reprezentacji. Zamiast korzystać z reprezentacji, wchodzi on w bezpośrednie interakcje ze środowiskiem. Samochód B działa na podstawie MKM, a zatem stanowi prosty system reprezentacyjny. Jaki jest jednak status samochodu A? Zauważmy przede wszystkim, że korzysta on ze swego rodzaju receptorów, jakimi są ulokowane po jego bokach wypustki. Wypustki te (1) współzmieniają się z bliskością krawędzi toru (wypustka porusza się zawsze wtedy, gdy krawędź znajduje się w określonej odległości od samochodu) oraz (2) zawdzięczają tej współzmienności swoją funkcję w pojeździe (sterują nim dzięki tej współzmienności). Czy samochód A korzysta jednak z reprezentacji? Możemy to pytanie sformułować też w następujący sposób: czy samochód A przypomina bardziej pozbawiony reprezentacji samochód C, czy wykorzystujący reprezentacje samochód B? Zgadzam się z diagnozą Ramseya (2007: 197–199), że sposób funkcjonowania samochodu A przypomina bardziej to, jak tor jest pokonywany przez (niereprezentacyjny) samochód C niż przez (reprezentacyjny) samochód B. Jak to ujmuje sam autor:

> Obydwa te procesy [to znaczy pokonywanie toru przez samochody A i C – P. G.] możemy wyjaśnić jako prostą interakcję między kołami samochodu a krawędzią toru. Ciąg przyczynowy w samochodzie A jest co prawda dłuższy. W wypadku tego pojazdu występuje więcej ogniw pośrednich między zbliżaniem się samochodu do krawędzi a zwrotem kół w przeciwną stronę. Nie ma jednak intuicyjnego sensu mówienie, że dołączona do kierownicy wypustka [...] czy którykolwiek z innych komponentów samochodu A pełni rolę reprezentacji w większym stopniu, niż odbijające się od krawędzi toru koła samochodu C. Samochód A także odbija się od ścian – tylko w nieco bardziej złożony sposób (Ramsey 2007: 197).

Sądzę, że powyższa diagnoza nie zmieni się, nawet kiedy zaczniemy modyfikować mechanizm mediacji przyczynowej w samochodzie A. Na przykład nawet wtedy, gdy wypustki zastąpimy jakąś formą fotoreceptorów systematycznie reagujących na zbliżanie się krawędzi toru, nie zmienimy statusu pojazdu. Zamienimy w ten sposób jeden rodzaj pośrednika przyczynowego na inny, jednak nie spra-

wimy, że pośredniczenie przyczynowe stanie się reprezentowaniem (por. Ramsey 2007). Sądzę także, że status samochodu A nie zmieni się (co stoi w kontrze do twierdzeń przedstawionych w: García, Calvo 2010) wtedy, gdy jego działanie zmodyfikujemy tak, by – zamiast odbijać się od krawędzi do krawędzi (jak samochód C) – poruszał się on płynnie środkiem toru (jak samochód B). Na przykład nie zmieni się on w system reprezentacyjny nawet wtedy, gdy wypustki po jego bokach będą na tyle długie, by zawsze dotykać krawędzi toru i w ten sposób „behawioralnie" upodobnić samochód A do samochodu C („upłynnić" jego działanie). O niereprezentacyjnym statusie samochodu A nie decyduje bowiem brak płynnego ruchu (zachowanie systemu jako całości), lecz natura mechanizmu, który za ten ruch odpowiada (wewnętrzna struktura systemu).

Bezpośrednia konfrontacja modeli mentalnych z (rzekomymi) reprezentacjami receptorowymi pozwala na podtrzymanie wyjściowej tezy, że tylko pierwsze, ale nie drugie zasługują na miano reprezentacji. Skąd dokładnie bierze się jednak ta różnica? Jaka jest jej podstawa? Intrygującą spekulację na temat jej źródeł przedstawia Ramsey (2007: 200–202). Wychodzi on od obserwacji, że naukowe pojęcia, odpowiednio, wewnętrznych modeli oraz wewnętrznych receptorów mają różne rodowody, to znaczy są inaczej zakorzenione w naszych przednaukowych sposobach konceptualizacji świata. Modele mentalne są odpowiednikiem pozamentalnych, przednaukowych reprezentacji ikonicznych, natomiast receptory stanowią odpowiednik pozamentalnych, przednaukowych reprezentacji indeksowych. Ramsey sugeruje, że różnice w statusie eksplanacyjnym (wewnętrznych) modeli i receptorów mają u swoich podstaw różnice zachodzące między odpowiadającymi im przednaukowymi pojęciami reprezentacji. Spekuluje on, że różnice między pozamentalnymi reprezentacjami indeksowymi a ikonicznymi nie dotyczą jedynie (jak sugerowaliśmy wcześniej) relacji nośnik–przedmiot, ale także sposobu funkcjonowania. Reprezentacje ikoniczne stanowią podstawę rozumowań surogatywnych i w ten sposób zastępują określone przedmioty dla swoich użytkowników. Według Ramseya rola (pozamentalnych) indeksów polega z kolei na *informowaniu* podmiotu o zajściu stanu rzeczy czy zdarzenia. Reprezentacje indeksowe *im-*

plikują lub *pociągają za sobą* (*entail*) zajście określonych okoliczności – właśnie w odniesieniu do współzmienności między nośnikiem a tymi okolicznościami. Interpretacja tego rodzaju reprezentacji wymaga zatem od interpretatora posiadania pewnych zdolności inferencyjnych, pozwalających wywnioskować stan przedmiotu reprezentacji ze stanu, w jakim znajduje się nośnik.

Wedle sugestii Ramseya na powyższym fakcie opiera się różnica w statusie eksplanacyjnym czysto mechanicznych odpowiedników reprezentacji ikonicznych (modeli mentalnych) oraz czysto mechanicznych odpowiedników reprezentacji indeksowych (receptorów). Jak zobaczyliśmy na przykładzie samochodu Cumminsa oraz Bezmyślnego Jana, rozumowania surogatywne mogą być wykonywane mechanicznie, bez udziału pełnoprawnego podmiotu intencjonalnego. To samo nie może być powiedziane o wnioskowaniu o przedmiocie reprezentacji na podstawie nośnika reprezentacji indeksowej: „bez umysłu wykonującego odpowiednie rozumowanie, relacja zależności [współzmienności – P. G.] może zostać jakoś użyta [...], jednak nie w celach reprezentacyjnych" (Ramsey 2007: 201). Reprezentacje indeksowe z konieczności wymagają udziału myślących podmiotów. Postulując wewnętrzne indeksy pozbawione takich interpretujących podmiotów, jesteśmy albo skazani na pomylenie reprezentowania z pośrednictwem przyczynowym, albo na popełnienie błędu homunkularnego (na przypisywanie komponentom mechanizmu zdolności inferencyjnych, których istnienia nie powinniśmy postulować na tym poziomie organizacji systemu poznawczego). Status eksplanacyjny wewnętrznych ikon i wewnętrznych indeksów będzie zatem zasadniczo różny (por. podobne rozważania zawarte w: Cummins, Poirier 2004).

e) Nie ma żadnej różnicy między modelami mentalnymi a receptorami
Natura problemu: W poprzednim punkcie polemizowałem z tezą głoszącą, że receptory wcale nie radzą sobie gorzej z wymogiem opisu zadań niż S-reprezentacje. Nie negowałem jednak samego założenia, że te ostatnie rzeczywiście różnią się jakoś od receptorów. Teraz chciałbym skupić się na zarzucie, zgodnie z którym sama dystynkcja

między receptorami a modelami mentalnymi (wewnętrznymi S-reprezentacjami) nie może być przeprowadzona.

Alex Morgan (2014) zwrócił uwagę na fakt, że przy bliższym spojrzeniu okazuje się, iż znakomita większość struktur klasyfikowanych przez Ramseya jako receptory może zostać równie naturalnie i zasadnie zaliczona do klasy S-reprezentacji. Jak stwierdza ten autor, „receptory *po prostu są* reprezentacjami strukturalnymi" (Morgan 2014: 236). W jego argumentacji kluczowe znaczenie ma obserwacja, że receptory nie tylko współzmieniają się z pewnymi okolicznościami zewnętrznymi, ale są z nimi także powiązane *relacją podobieństwa strukturalnego* (Morgan pisze konkretnie o homomorfizmie). Spróbujmy zilustrować tę myśl. Weźmy pod uwagę magnetosomy w bakteriach morskich, wymienione wcześniej jako klarowny przykład receptorów („reprezentacji" receptorowych). Opisując funkcję pełnioną przez magnetosomy, na ogół zwraca się uwagę na znaczenie faktu, iż położenie zawartych w nich kryształków magnetytu systematycznie współzmienia się z położeniem północy magnetycznej względem bakterii. Jednak podkreślając rolę współzmienności, nie dostrzega się innego faktu. Otóż między potencjalnymi położeniami przestrzennymi kryształków zawartych w magnetosomie (możemy je sobie wyobrazić jako możliwe położenia strzałki w bakteryjnym „kompasie") a możliwymi położeniami północy magnetycznej względem bakterii zachodzi także strukturalne podobieństwo. Można przypisać poszczególne możliwe położenia kryształków w magnetosomie możliwym położeniom północy magnetycznej względem bakterii w taki sposób, że układ czy wzór tych pierwszych (struktura możliwych kierunków, w jakie przechyla się strzałka bakteryjnego „kompasu") będzie strukturalnie odzwierciedlać układ czy wzór tych drugich (strukturę możliwych położeń północy względem bakterii). Wydaje się więc, że to, co pierwotnie jawiło się jako klarowny przykład receptora, okazuje się w istocie formą S-reprezentacji.

Aby jeszcze inaczej zilustrować ten sposób myślenia, przyjrzymy się przedstawionemu wyżej zestawieniu „receptorowej" i „S-reprezentacyjnej" wersji samochodu Cumminsa, czyli samochodom A i B. Ta pierwsza – wyposażona w wypustki stykające się (współ-

zmieniające) z krawędziami toru – wydaje się początkowo działać na zupełnie innej zasadzie niż druga, wykorzystująca wewnętrzną mapę. Jednak rozumowanie Morgana pokazuje, a w każdym razie wydaje się pokazywać, że i w tym przypadku różnicę można określić jako pozorną. W samochodzie A (tym „receptorowym") poziom wychylenia wypustki jest proporcjonalny do bliskości przestrzennej krawędzi toru względem samochodu. Na przykład im bliżej pojazdu znajduje się prawa krawędź toru, tym mocniej wychyla się (w lewo) prawa wypustka. Struktura relacji między możliwymi poziomami wychylenia wypustek odzwierciedla strukturę możliwych odległości przestrzennych między samochodem a krawędziami toru. Raz jeszcze można uznać, że rzekomy receptor okazuje się w istocie wewnętrznym modelem. Wbrew pozorom samochody A i B wcale się nie różnią.

Według Morgana tego samego rodzaju rozumowanie można przeprowadzić dla dowolnej wewnętrznej struktury będącej rzekomo receptorem – nawet takiej, która może znajdować się tylko w dwóch stanach (na przykład: aktywacja/brak aktywacji). Sama dystynkcja między S-reprezentacjami a receptorami jest zatem bezpodstawna. Nie ma receptora, który nie byłby S-reprezentacją, czyli modelem. Z konstatacji tej można potencjalnie wyciągnąć dwa alternatywne, wykluczające się wnioski. Z jednej strony można stwierdzić, że w świetle takiej diagnozy wszelkie receptory okazują się w istocie S-reprezentacjami, a zatem wbrew pozorom spełniają one wymóg opisu zadań. Z drugiej – można wyciągnąć wniosek, że skoro rzekome wewnętrzne modele w istocie nie różnią się od struktur klasyfikowanych jako receptory, to nie spełniają one – podobnie jak receptory – wymogu opisu zadań. Wydaje się, że sam Morgan skłania się ku akceptacji tej ostatniej opcji[32]. Zamiast argumentować za

[32] Stanowisko Morgana (2014) jest w istocie nieco bardziej subtelne. Argumentuje on, że wewnętrzne S-reprezentacje nie zasługują na miano *mentalnych*. Jednak autor ten wydaje się przyjmować stosunkowo restrykcyjne kryteria „bycia modelem mentalnym", w świetle których kwalifikacja czegoś jako „mentalnego" wymaga, aby było ono jakoś zaangażowane w realizowanie zaawansowanych, wyższych zdolności poznawczych (związanych z dokonywanym *off-line* myśleniem kontrfaktycznym czy wyobraźnią). W pracy tej nie podzielam takiej

jedną z tych dwóch alternatywnych konkluzji, chcę teraz raczej uderzyć w argumentację tego autora, podważając konkluzywność jego ataku na dystynkcję między receptorami a S-reprezentacjami.

Odpowiedź: Problem stawiany przez Morgana jest być może najpoważniejszy ze wszystkich, jakie są w tej sekcji podejmowane. Dystynkcja między receptorami a S-reprezentacjami, która początkowo wydawała się intuicyjnie jasna i teoretycznie uzasadniona, okazuje się pod wpływem argumentacji tego autora zupełnie nieoczywista. Jednakże mimo niewątpliwej siły argumentu Morgana sądzę, że może on być z powodzeniem odparty. Wewnętrzne ikony różnią się od wewnętrznych indeksów. Zachodzą między nimi trzy znaczące teoretycznie różnice, których ten autor nie dostrzega.

Po pierwsze, należy wyraźnie odróżnić (1) relacje, jakie zachodzą między nośnikiem a przedmiotem reprezentacji od (2) relacji zachodzących między nośnikiem a przedmiotem reprezentacji, które są zarazem *relewantne dla funkcjonowania nośnika*. Argumentacja Morgana (2014) wydaje się pomijać to zasadnicze rozróżnienie. Autor ten pokazuje, że między receptorami a tym, z czym one się współzmieniają, zachodzi – oprócz samej kowariancji – relacja podobieństwa strukturalnego. Nie formułuje on jednak argumentu za tym, że relacja podobieństwa jest *relewantna* dla funkcjonowania receptorów. Tymczasem tu właśnie leży pierwsza ważna różnica między S-reprezentacjami a receptorami. Jak już zostało pokazane w poprzedniej sekcji, poprawne funkcjonowanie S-reprezentacji czy modeli w szerszych mechanizmach systematycznie zależy od relacji podobieństwa między samą reprezentacją (jej nośnikiem) a tym, co reprezentowane. Podobieństwo to relacja relewantna dla funkcjonowania tego typu reprezentacji. Jednak podobieństwo nie jest relewanantne dla funkcjonowania receptorów. Co prawda w pewnym sensie receptory są podobne strukturalnie do określonych okoliczności zewnętrznych (tu Morgan ma rację), jednak to nie podobień-

restrykcyjności w posługiwaniu się kategorią „mentalny". Przyjmuję, że nawet subosobowe mechanizmy czy komponenty realizujące niższe albo proste funkcje poznawcze mogą być uznane za „mentalne". Prosta aktywność umysłowa czy poznawcza to nadal aktywność umysłowa/poznawcza.

stwo, lecz *współzmienność* z tymi okolicznościami decyduje o ich poprawnym funkcjonowaniu w ramach szerszej całości (mechanizmu). Ich poprawne funkcjonowanie zależy od zachodzenia współzmienności, a nie podobieństwa: obie relacje zachodzą, lecz to ta pierwsza jest relewantna. Na przykład teoretycznie moglibyśmy zaprojektować magnetosom tak, by jego stany współzmieniały się z okolicznościami środowiskowymi, jednak w sposób arbitralny, nierespektujący czy odzwierciedlający tych okoliczności. W takiej sytuacji każde możliwe położenie „strzałki" (kryształków magnetytu) magnetosomu–kompasu odpowiadałby jakiemuś możliwemu położeniu północy magnetycznej, lecz nie w taki sposób, aby struktura możliwych położeń przestrzennych „strzałki" odzwierciedlała strukturę możliwych położeń przestrzennych północy magnetycznej. Tak zmodyfikowany, „pozbawiony" podobieństwa strukturalnego magnetosom mógłby nadal działać poprawnie, gdybyśmy jednocześnie: (1) utrzymali zachodzenie systematycznej współzmienności między stanami magnetosomu a okolicznościami zewnętrznymi (to oznacza, że stany magnetosomu są arbitralnie przypisane do położeń północy magnetycznej, jednak nadal systematycznie się z nimi współzmieniają), (2) odpowiednio zmodyfikowali powiązania między stanami magetosomu a reakcjami behawioralnymi bakterii (tak, by jej zachowania pozostały adaptywnie dostosowane do położenia północy magnetycznej). Pokazuje to, że magnetosom jest bakterii „potrzebny", o ile współzmienia się on z okolicznościami zewnętrznymi. To współzmienność okazuje się relewantna dla jego działania (funkcji) w mechanizmie. Podobieństwo strukturalne, choć zachodzi, jest tu w zasadzie „epifenomenalne"[33]. Uważam, że tę konkluzję można zgeneralizować tak, by dotyczyła wszelkich receptorów.

[33] Twierdzę, że w przypadku S-reprezentacji czy modeli takie uniezależnienie funkcji pełnionej przez nośnik reprezentacji od zachodzenia relacji podobieństwa jest niemożliwe. S-reprezentacje z konieczności nie mogą pełnić swojej funkcji inaczej, niż tylko na podstawie podobieństwa zachodzącego między nośnikiem a przedmiotem reprezentacji. Mapa w samochodzie Cumminsa albo diagram, z którego korzysta Bezmyślny Jan, nie mogłyby realizować swoich funkcji inaczej, niż tylko odzwierciedlając strukturę przedmiotu reprezentacji.

Po drugie, receptory w zasadzie *ex definitione* nie mogą działać w oderwaniu[34] od tego, z czym się współzmieniają. Nie mogą realizować swojej funkcji pod nieobecność przedmiotu reprezentacji. Jest bowiem prawdą pojęciową, że receptory nie mogą *współzmieniać* się z okolicznościami, które nie *współwystępują* z ich aktywnością. Receptor nie może funkcjonować w trybie *off-line*[35]. Tymczasem wewnętrzne S-reprezentacje mają to do siebie, że mogą realizować swoją funkcję w oderwaniu od przedmiotu reprezentacji, również pod całkowitą nieobecność tego przedmiotu. Kategoria „oderwania" modeli mentalnych zostanie omówiona w punkcie (f) tej sekcji, a póki co poprzestańmy na tej ogólnej uwadze.

Po trzecie wreszcie, należy zauważyć, że podobieństwo strukturalne między nośnikiem a przedmiotem reprezentacji może być rozumiane na dwa różne sposoby. Z jednej strony możemy to podobieństwo rozumieć w takim sensie, że: (1) zarówno nośnik, jak i przedmiot reprezentacji mogą znajdować się w dowolnym momencie *t* w jednym z szeregu alternatywnych, potencjalnych stanów; (2) potencjalne stany zarówno nośnika, jak i przedmiotu układają się w pewne struktury (można je uporządkować według pewnych relacji); (3) zachodzi podobieństwo drugiego rzędu między strukturą potencjalnych stanów nośnika a strukturą potencjalnych stanów przedmiotu reprezentacji. Nazwijmy taką relację „podobieństwem strukturalnym typu I". Z drugiej strony podobieństwo między no-

[34] Precyzyjniej: nie mogą działać w sposób słabo lub mocno oderwany od przedmiotu reprezentacji. Te różne gradacje oderwania reprezentacji od jej przedmiotu scharakteryzuję w kolejnym punkcie – (f).

[35] Morgan (2014) powołuje się na badania pokazujące, iż neurony w korze wzrokowej, uznawane często za detektory cech (czyli za pewną formę receptorów), są aktywne także podczas generowania wyobrażeń wzrokowych. Jak się jednak wydaje, z faktu tego nie wynika – wbrew sugestiom wymienionego autora – że w korze wzrokowej istnieją oderwane reprezentacje receptorowe (receptory działające *off-line*). Powinniśmy raczej uznać, że struktury neuronalne klasyfikowane jako proste detektory cech być może tak naprawdę realizują funkcję modelu. Jak zauważa Sprevak, „detektory twarzy czy krawędzi są rozumiane [we współczesnej neuronauce – P. G.] nie jako detektory w izolacji, lecz jako składowe szerszego modelu świata, [...] pełniące w rozumowaniach rolę użytecznych surogatów twarzy i krawędzi" (2011: 765).

śnikiem a przedmiotem reprezentacji zachodzić może nie tyle na poziomie struktur *wszystkich potencjalnych stanów* nośnika i przedmiotu, lecz *wewnątrz pojedynczych stanów*. Co innego bowiem powiedzieć, że struktura potencjalnych stanów nośnika odzwierciedla strukturę potencjalnych stanów przedmiotu, a co innego stwierdzić, że same pojedyncze stany nośnika odzwierciedlają „wewnętrznie" pojedyncze stany przedmiotu. W tym ostatnim przypadku to sam pojedynczy stan, w jakim nośnik znajduje się w pewnym momencie *t*, traktujemy jako strukturę – układ elementów powiązanych relacjami – i stwierdzamy, że odzwierciedla ona strukturę stanu, w jakim znalazł się przedmiot reprezentacji. Nazwijmy podobieństwo w takim znaczeniu „podobieństwem typu II".

Po co nam jednak rozróżnienie między podobieństwami typu I oraz II? Zauważmy, że kiedy Morgan twierdzi, iż receptory odzwierciedlają strukturalnie pewne zewnętrzne okoliczności, wydaje się, że ma on na myśli *podobieństwo typu I*. Da się to zilustrować za pomocą nieskomplikowanego przykładu. Możemy wyobrazić sobie prosty receptor-czujnik: żarówkę, która świeci tym jaśniej, im bliżej niej znajduje się pewien kot. Mamy tu do czynienia z podobieństwem typu I: struktura możliwych poziomów jasności emitowanego przez żarówkę światła (to znaczy struktura potencjalnych stanów żarówki, wyróżnionych oraz uporządkowanych według jasności emitowanego światła) odzwierciedla strukturę możliwych odległości przestrzennych między kotem a żarówką (to znaczy strukturę możliwych położeń przestrzennych kota uporządkowanych według jego oddalenia od żarówki). Jednak nie mamy tu do czynienia z *podobieństwem typu II*. Każdy pojedynczy stan, w jakim żarówka znajduje się w danym momencie – emitowanie światła o określonej jasności – jest „atomowy". Nie składa się on z elementów połączonych relacjami w taki sposób, by odzwierciedlać strukturę obiektu (kota). Kiedy jednak mówię w tej pracy o modelach czy S-reprezentacjach i podkreślam wagę, jaką w ich funkcjonowaniu spełnia podobieństwo strukturalne między nośnikiem a przedmiotem reprezentacji, mam na myśli właśnie podobieństwo typu II. Nie twierdzę co prawda, że w modelach nigdy nie mamy do czynienia z podobieństwem typu I. Jednak to, co jest istotne czy relewantne dla aktywności mo-

deli, to właśnie podobieństwo typu II. Klarowny przykład stanowi tu diagram, z jakiego korzysta Bezmyślny Jan. W funkcjonowaniu tego diagramu istotny jest jego wewnętrzny układ przestrzenny oraz podobieństwo tego układu do wzoru relacji pokrewieństwa zachodzących między członkami rodziny. Nie chodzi o podobieństwo między strukturą możliwych „stanów" diagramu a strukturą możliwych „stanów" rodziny. Chodzi raczej o to, że pojedynczy „stan" diagramu – ułożenie jego elementów (imion i strzałek) na jeden z możliwych sposobów – odzwierciedla pojedynczy „stan" rodziny (określoną, zachodzącą w niej strukturę pokrewieństwa). Tu właśnie przebiega kolejna linia demarkacyjna oddzielająca receptory i modele. W tych pierwszych mamy co prawda do czynienia z podobieństwem typu I, jednak nie zachodzi podobieństwo typu II. Z kolei w tych drugich zasadniczą rolę odgrywa (także) podobieństwo typu II[36].

[36] Do twierdzenia tego należy dodać pewien ważny aneks. Wyobraźmy sobie S-reprezentację, która jest całkowicie oparta na strukturze diachronicznej (temporalnej), a nie synchronicznej (na przykład przestrzennej). Nośnik tej S-reprezentacji ewoluuje w czasie, to znaczy znajduje się w odmiennych stanach (lub pozostaje w tych samych) w następujących po sobie momentach. Wzorzec zmian, jakim on podlega, ma odzwierciedlać wzorzec zmian, jakim podlega pewien przedmiot reprezentacji. Jednak w każdym pojedynczym momencie *t* nośnik jest niezłożony czy „atomowy"; to znaczy, że jego struktura wewnętrzna w dowolnym momencie *t* niczego nie odzwierciedla (w istocie nie można tu nawet mówić o żadnej strukturze). Możemy sobie taką S-reprezentację wyobrazić jako żarówkę ustawicznie zmieniającą jasność emitowanego światła w taki sposób, by symulować fluktuacje odległości przestrzennej między nią samą a jakimś kotem. W żadnym pojedynczym momencie żarówka nie odzwierciedla niczego, nie posiada żadnej struktury wewnętrznej; po prostu świeci z określoną jasnością. Czy w takim przypadku między nośnikiem a przedmiotem reprezentacji zachodzi jedynie podobieństwo typu I? Nie. Istota tego przykładu polega na tym, że *temporalny* układ czy wzorzec zmian, jakim podlega nośnik (żarówka) odzwierciedla *temporalny* wzorzec zmian, jakim podlega to, co reprezentowane (położenie kota). Aby zrozumieć istotę tego przykładu, potencjalne „stany" nośnika powinniśmy scharakteryzować jako interwały czasowe, od t_1 do t_n, w trakcie których nośnik podlega określonym zmianom. *Takie* stany nośnika są zatem „wewnętrznie" złożone – nie przestrzennie, lecz temporalnie. Stany te są zawsze strukturą relacji następstwa czasowego, jaka zachodzi między poziomami jasności światła emitowanego w trakcie interwału od t_1 do t_n. Przy takim ujęciu powinno być jasne, że mamy tu wbrew pozorom do czynienia z podobieństwem typu II: struktura temporalna pojedynczego stanu nośnika

Podsumowując: wbrew twierdzeniu Morgana zachodzi szereg istotnych różnic między modelami a receptorami. Ta niezwykle ważna z punktu widzenia tej pracy dystynkcja może zostać utrzymana.

f) Reprezentacje postulowane przez teorię MKM nie są (wystarczająco) oderwane od swoich przedmiotów

Natura problemu: Niektórzy współcześni autorzy bronią idei, że jedną z ważnych, a nawet konstytutywnych własności reprezentacji jest to, iż mogą być one *oderwane* (*decoupled*, *detatched*) od swoich przedmiotów (Clark, Toribio 1994; Gardenförs 1996; Clark 1997; Grush 1997; Haugeland 1998; Clark, Grush 1999; Chemero 2009: 55–65; García, Calvo 2010). Zasadniczą cechę każdej reprezentacji stanowi więc fakt, iż może ona być wykorzystywana pod nieobecność tego, co reprezentowane – wtedy, gdy przedmiot reprezentacji nie jest, jak to ujmuje Haugeland (1998), „reliabilnie obecny" (*reliably present*) dla systemu czy mechanizmu posługującego się reprezentacją. Jak stwierdza Andy Clark, system czy mechanizm wykorzystujący reprezentacje powinien wykazywać „zdolność do wykorzystywania wewnętrznych stanów w celu przewodzenia swoim działaniem pod nieobecność [reprezentowanych – P. G.] okoliczności środowiskowych" (Clark 1997: 144). Tę myśl dobrze ilustruje także proponowane przez Ricka Grusha (1997) rozróżnienie na prezentację i reprezentację. System wykorzystujący prezentację pewnego przedmiotu działa na podstawie bezpośredniego przyczynowego z nim sprzężenia. Na przykład taki „prezentacyjny" system może działać dzięki temu, że ma aparat sensoryczny, który systematycznie wchodzi w przyczynowe interakcje z generowanym przez pewien zewnętrzny przedmiot sygnałem chemicznym. Z kolei z systemem wykorzystującym reprezentacje mamy wedle Grusha do czynienia dopiero wtedy, gdy dysponuje on zdolnością do używania pewnych wewnętrznych struktur (reprezentacji) zamiast wchodzenia w bezpośrednie interakcje z określonym przedmiotem. Reprezentacje są wykorzystywane jako stosowane *off-line* (poza kontekstem interakcji),

---

odzwierciedla (powinna docelowo odzwierciedlać) strukturę temporalną stanu przedmiotu reprezentacji.

„kontrfaktyczne prezentacje" przedmiotu reprezentacji, pozwalające przewidzieć jego przyszłe zachowanie albo zaledwie możliwe czy potencjalne skutki określonych działań zorientowanych na ten przedmiot. Dowolna struktura wewnętrzna może stanowić reprezentację dopiero wtedy, gdy jest ona (przynajmniej potencjalnie) oderwana od bezpośrednich interakcji z przedmiotem.

Powstaje wątpliwość: czy postulowane przeze mnie modele mentalne są reprezentacjami oderwanymi w opisanym wyżej sensie? Czy teoria mechanizmów reprezentacyjnych jako wyposażonych w konsumowane modele dopuszcza możliwość, aby reprezentacje były oderwane od swoich przedmiotów? Wykorzystywany tu często przykład samochodu Cumminsa wydaje się sugerować odpowiedź negatywną. Zawarta w tym pojeździe mapa toru nie może przecież spełniać swojej roli poza kontekstem bezpośrednich, dokonywanych *on-line* interakcji na linii samochód–tor.

Odpowiedź: Na powyższe wątpliwości można by zapewne odpowiedzieć, odrzucając warunek, zgodnie z którym reprezentacje powinny być oderwane od swoich przedmiotów. Sądzę jednak, iż taki zabieg nie jest w ogóle potrzebny, ponieważ postulowane przeze mnie modele wewnętrzne są reprezentacjami oderwanymi.

Zacznijmy od zwrócenia uwagi na fakt, że omówiona wyżej rekonstrukcja przedstawia dość restrykcyjne pojmowanie „oderwania" reprezentacji. Sugeruje ona bowiem, że reprezentacja może zostać określona jako oderwana jedynie wtedy, gdy jej przedmiot jest całkowicie nieobecny dla systemu czy mechanizmu z tej reprezentacji korzystającego. Jednakże niektórzy filozofowie zajmujący się kategorią oderwania odrzucili tak mocny wymóg (Clark 1997; Clark, Grush 1999). Zdecydowali się oni dopuścić, że reprezentacja może być zakwalifikowana jako oderwana nawet wtedy, gdy możliwość jej zastosowania ogranicza się do sytuacji, w których dochodzi do bezpośrednich interakcji szerszego systemu (korzystającego z owej reprezentacji) z przedmiotem reprezentacji. Taka modyfikacja była motywowana chęcią obrony tezy, że zaproponowana przez Grusha teoria emulacji postuluje istnienie oderwanych reprezentacji (Clark

1997; Clark, Grush 1999; por.: Chemero 2009: 60–65; García, Calvo 2010).

Naturze emulatorów jako formy modeli mentalnych zostanie poświęcone więcej miejsca w dalszej części tego rozdziału (w podrozdziale 4.3). W tej chwili powiedzmy tylko bardzo ogólnie, że emulatory to postulowane przez niektórych kognitywistów wewnętrzne struktury, które miałyby pełnić w systemie poznawczym rolę zastępowania czegoś innego (a zatem rolę reprezentacji), jednak (1) dla swojego działania wymagają ustawicznego dopływu sygnałów z przedmiotu reprezentacji (na przykład sygnałów proprioceptywnych z ciała); (2) realizują swoją funkcję tylko w sytuacji, w której szerszy system czy mechanizm wchodzi w interakcje z reprezentowanym przedmiotem. Emulatory wydają się intuicyjnie pełnić funkcję reprezentacji – zastępują coś dla systemu, są wewnętrznymi „surogatami" – choć czynią to tylko w kontekście dokonywanych *on-line* interakcji z przedmiotem reprezentacji. Co jednak bardzo istotne, mogą one wykonywać swoją funkcję także w ramach krótkich interwałów czasowych, w trakcie których nie docierają do nich sygnały z przedmiotu reprezentacji. Wydaje się zatem uzasadnione twierdzić, że emulatory są reprezentacjami w jakimś sensie oderwanymi. Aby nadać emulatorom status reprezentacji oderwanych, niektóry autorzy przyjęli, że oderwanie reprezentacji jest *stopniowalne* (Clark 1997; Clark, Grush 1999; por. Chemero 2009: 60–65). Emulatory miałyby z takiej perspektywy stanowić przykład reprezentacji słabo oderwanych od interakcji ze swym przedmiotem. Stanowiłyby one jeden kraniec skali „oderwania", którego drugi kraniec miałyby stanowić struktury pełniące rolę reprezentacji pod całkowitą nieobecność ich przedmiotu.

Opisaną wyżej strategię opartą na stopniowaniu oderwania reprezentacji można zgeneralizować, tak aby obejmowała nie tylko emulatory, ale wszelkie możliwe rodzaje modeli mentalnych. Proponuję wyróżnić trzy stopnie oderwania reprezentacji od jej przedmiotu:

1. Reprezentacja jest *minimalnie oderwana*, jeśli nie zachodzi przyczynowa interakcja między konsumentem a przedmiotem tej reprezentacji.
2. Reprezentacja jest *słabo oderwana*, jeśli (1) nie zachodzi przyczynowa interakcja między konsumentem a przedmiotem tej reprezentacji, a także (2) nie zachodzi przyczynowa interakcja między nośnikiem a przedmiotem tej reprezentacji.
3. Reprezentacja jest *mocno oderwana*, jeśli korzystający z niej system lub mechanizm nie wchodzi w interakcje przyczynowe z jej przedmiotem.

Zilustrujmy te trzy kategorie za pomocą poglądowych przykładów:

Ad 1. Wyobraźmy sobie pewien wariant samochodu Cumminsa. Przemierza on S-kształtny tor, posługując się wewnętrzną mapą, jednak mapa ta nie preegzystuje w systemie, lecz jest na bieżąco tworzona w wyniku interakcji przyczynowych z samym torem. Od tablicy znajdującej się wewnątrz samochodu odchodzą dwie wypustki – podobne do posiadanych przez samochód A z przykładu opisanego w punkcie (d) tej sekcji – które stykają się z krawędziami toru. Wypustki te są sprzężone z układem, który na bieżąco ryje w tablicy kształt odpowiadający kształtowi toru[37]. Z tej tworzonej ustawicznie mapy korzysta ster (konsument reprezentacji) zawiadujący ruchami pojazdu. Dlaczego twierdzę, że taka wewnętrzna mapa w ogóle zasługuje na miano reprezentacji oderwanej? Otóż pamiętajmy, że z bronionej tu perspektywy, o statusie czegoś jako reprezentacji decyduje istnienie triady nośnik–przedmiot–konsument. Prowadzone dotychczas w literaturze dyskusje dotyczące oderwanych reprezentacji koncentrowały się na położeniu przyczynowym ich nośników. Sądzę jednak, iż można także utrzymywać, że reprezentacja jest oderwana ze względu na położenie przyczynowe jej *konsumenta*. Nawet jeśli nośnik jest konstruowany na podstawie „sygnałów"

[37] Taki mechanizm odpowiadający za wytworzenie nośnika reprezentacji możemy nazwać, czerpiąc z terminologii Millikan (2002), producentem reprezentacji.

płynących ze środowiska zewnętrznego, to nadal zastępuje on przedmiot reprezentacji dla konsumenta. Nośnik bowiem kompensuje konsumentowi fakt, że ten nie jest bezpośrednio powiązany przyczynowo z przedmiotem reprezentacji. Innymi słowy, przedmiot reprezentacji nie jest bezpośrednio „dostępny" przyczynowo dla konsumenta. Dlatego też mamy tu do czynienia z formą reprezentacji oderwanej, choć oderwanej minimalnie.

Ad 2. Przykładem reprezentacji słabo oderwanej jest mapa toru, z której korzysta samochód Cumminsa w wyjściowej wersji (czyli ten wyposażony w preegzystującą mapę). W pojeździe tym ani nośnik, ani konsument reprezentacji nie wchodzą w interakcje przyczynowe z jej przedmiotem. Model, którym posługuje się ten pojazd, jest zatem oderwany od przedmiotu reprezentacji w większym stopniu, niż w przypadku opisanym powyżej. Oderwanie nadal nie jest jednak pełne, ponieważ reprezentacja ta może spełniać swoją funkcję tylko w sytuacji, w której szerszy system (samochód jako całość) wchodzi w interakcje z jej przedmiotem. Wewnętrzna mapa jest bowiem funkcjonalnie bezużyteczna, dopóki samochód nie zacznie przemierzać toru.

Ad 3. Aby zachować ciągłość z przykładami wymienionymi w powyższych punktach, spróbuję raz jeszcze odpowiednio zmodyfikować samochód Cumminsa. Wyobraźmy sobie bardzo złożoną wersję tego pojazdu. Wyposażono go nie tyle w model toru, co *model samego siebie poruszającego się po torze*. Ten wewnętrzny model przyjmuje postać miniatury samochodu, która wykorzystując miniaturę mapy, przemierza miniaturową wersję toru (jest to po prostu wewnętrzna, ruchoma makieta pojazdu). Taki miniaturowy samochód może być ustawiany w miniaturowych torach o różnym kształcie. Można w nim też umieszczać różne warianty wewnętrznych map, za pomocą których porusza się on po tych miniaturowych torach. Taka miniatura stanowi dla większego samochodu manipulowany model samego siebie, dzięki któremu może on „weryfikować", jak sam by się zachowywał w różnych kontrfaktycznych sytuacjach, to znaczy

przy różnych zestawieniach mapa–tor (do czego doprowadziłoby zastosowanie mapy o takim a takim kształcie w celu przemierzenia toru o takim a takim kształcie). Możemy sobie wyobrazić, że owo wewnętrzne symulowanie różnych alternatywnych scenariuszy dokonuje się poza kontekstem interakcji z rzeczywistym torem. Tego rodzaju symulacja może chociażby służyć preselekcji odpowiedniej, zapewniającej sukces nawigacyjny mapy, zanim jeszcze dojdzie do właściwej interakcji. Stojąc na przykład przed wyzwaniem polegającym na pokonaniu w przyszłości toru o kształcie litery „S" – w ramach procesu symulacji pojazd może wybrać tę spośród alternatywnych map, która wewnętrznej miniaturze pozwoliła na najsprawniejsze pokonanie miniatury S-kształtnego toru[38]. Reprezentacja, którą posługuje się tak zmodyfikowany samochód Cumminsa, całkowicie spełnia postulat Haugelanda, zgodnie z którym przedmiot reprezentacji nie jest „reliabilnie dostępny" dla szerszego systemu. Wewnętrzna miniatura pełni rolę reprezentacji w sytuacji, w której większy, zawierający ją system w ogóle nie wchodzi w interakcje przyczynowe z jej przedmiotem. Jest to zatem reprezentacja mocno oderwana od swojego przedmiotu.

Podsumowując, wydaje się, że MKM są wyposażone w reprezentacje oderwane. W zależności od przyczynowego położenia konsumenta, nośnika reprezentacji albo całego zawierającego reprezentację systemu czy mechanizmu, modele mentalne mogą być minimalnie, słabo lub mocno oderwane.

[38] Przyjmijmy, że te miniaturowe mapy są najpierw generowane losowo, a następnie kolejno testowane w trybie *off-line*. Na pewnym etapie jest selekcjonowany ten z przetestowanych wariantów, który w ramach symulacji zapewnił miniaturze pojazdu najsprawniejsze pokonanie miniaturowego toru. Na podstawie tak wybranej miniatury tworzy się większą mapę, którą właściwy samochód wykorzysta do pokonania właściwego toru. (Komponent generujący właściwą mapę na podstawie wyselekcjonowanej miniaturowej mapy stanowiłby tu konsumenta reprezentacji, który w swoim działaniu wykorzystuje podobieństwo między zachowaniem wewnętrznego modelu a zachowaniem pojazdu jako całości).

g) MKM mogą nie wyczerpywać klasy mechanizmów reprezentacyjnych
Natura problemu: Broniona tu teza – mówiąca, że mechanizmy reprezentacyjne to mechanizmy korzystające z wewnętrznych modeli – jest stosunkowo silna. Przy określonym odczytaniu głosi ona, że MKM wyczerpują klasę mechanizmów reprezentacyjnych. Czy nie jest to jednak zbyt mocne twierdzenie? Czy nie mogą istnieć inne rodzaje mechanizmów, które także zasługują na takie miano? Odrzuciłem do tej pory niektóre potencjalne rozwiązania problemu mechanizmów reprezentacyjnych, takie, które mogłyby odwoływać się do reprezentacji rozumianych jako receptory, przewodniki działań oraz jako „ukryte" struktury odpowiadające za własności dyspozycyjne systemów poznawczych. Nie wydaje się jednak, by ta lista wyczerpywała wszystkie możliwe alternatywy dla koncepcji MKM. Skoro nie odrzuciłem zatem wszystkich alternatyw, jak mogę zawężać kategorię mechanizmów reprezentacyjnych jedynie do takich, które są wyposażone w konsumowane modele? Czy taka koncepcja nie prowadzi do swoistego „modelowego szowinizmu"?

Odpowiedź: Odróżnijmy dwie możliwe interpretacje głównej tezy tej książki:

> **(I1)** Dla dowolnego mechanizmu poznawczego *M*, jeśli *M* stanowi MKM, to *M* stanowi mechanizm reprezentacyjny.
> **(I2)** Dla dowolnego mechanizmu poznawczego *M*, *M* stanowi mechanizm reprezentacyjny wtedy i tylko wtedy, gdy *M* stanowi MKM.

Teza (I1) jest słabsza i dopuszcza możliwość istnienia mechanizmów reprezentacyjnych, które nie są MKM. Teza (I2) jest mocniejsza i nie dopuszcza takiej możliwości.

Sporą część tego rozdziału poświęciłem argumentowaniu za tym, że MKM zasługują na status mechanizmów reprezentacyjnych. Utrzymuję więc, że (I1) jest prawdziwa. Postać gramatyczna mojej tezy głównej („mechanizmy reprezentacyjne to mechanizmy korzystające z wewnętrznych modeli") wyraźnie sugeruje jednak, że chcę także zaakceptować jej mocną interpretację, czyli (I2). Sądzę, że

prawdziwość (I2) jest bardzo prawdopodobna. Akceptuję zatem tę interpretację tezy głównej, choć – co należy dobitnie podkreślić – jedynie hipotetycznie (prowizorycznie). W poprzednim rozdziale pokazałem, że nawet mocno zakorzenione w praktyce eksplanacyjnej kognitywistów sposoby pojmowania reprezentacji nie przechodzą testu opartego na wymogu opisu zadań. Przyznaję jednak, że całkowicie konkluzywne uzasadnienie (I2) wymagałoby przeprowadzenia wyczerpującej krytyki wszelkich wiarygodnych, alternatywnych dla koncepcji MKM teorii mechanizmów reprezentacyjnych[39]. Ponadto,

[39] Warto zauważyć chociażby, że Ramsey (2007: 67–117) twierdzi, iż w kognitywistyce są wykorzystywane dwa pojęcia reprezentacji spełniające wymóg opisu zadań. Oprócz pojęcia S-reprezentacji jest to według niego pojęcie IO-reprezentacji (reprezentacji typu wejście–wyjście). Przypomnijmy: posługiwanie się pojęciem IO-reprezentacji polega na przypisywaniu treści intencjonalnych symbolom biorącym udział w procesach obliczeniowych realizowanych przez pewien system czy mechanizm obliczeniowy (por. sekcja 3.2.2). Czy możemy zatem stworzyć koncepcję mechanizmów reprezentacyjnych jako mechanizmów korzystających z IO-reprezentacji? Choć werdykt Ramseya w sprawie eksplanacyjnego statusu IO-reprezentacji jest pozytywny, to wątpliwym pozostaje, czy rola reprezentacji tego rodzaju będzie rzeczywiście zgodna z wymogami nakładanymi na eksplanacyjnie wartościowe reprezentacje przez mechanicyzm. Dlaczego? Przypisywanie treści symbolom biorącym udział w obliczeniach ma z pewnością istotną wartość heurystyczną i stanowi ważne wsparcie epistemiczne dla badaczy. Nie jest jednak pewne, na jakiej podstawie mamy uznać, że treść ta (a nie jedynie syntaktyczne czy formalne własności przetwarzanych symboli) określa czy determinuje rodzaj roli funkcjonalnej pełnionej przez symbole w ramach mechanizmu. Nie ma dobrych racji, by sądzić, że procesy obliczeniowe są metafizycznie indywiduowane na podstawie własności semantycznych czy intencjonalnych (Piccinini 2008; Miłkowski 2013). IO-reprezentacje są zatem narzędziami heurystycznymi, a nie komponentami, na których działaniu opiera się funkcjonowanie mechanizmów. Możemy ten zarzut sformułować także w następujący sposób: bycie IO-reprezentacją jest konstytutywnie (metafizycznie) pochodne względem praktyk epistemicznych podmiotów, które zajmują się konstruowaniem i badaniem systemów obliczeniowych (na przykład programistów albo specjalistów od sztucznej inteligencji). Prowadzi to do wniosku, że pojęcie IO-reprezentacji nie spełnia wymogów nakładanych przez mechanistyczny model wyjaśniana. Jak bowiem zaznaczyłem wcześniej (sekcja 3.1.2), zgodnie z wymogami stawianymi przez mechanicyzm reprezentacje powinny być realnymi, przyczynowo aktywnymi komponentami mechanizmów, a nie czysto instrumentalnymi konstruktami. Mechanizm reprezentacyjny powinien zachować swój status nawet pod nieobecność jakichkolwiek podmiotów, które

jak już wcześniej zaznaczyłem (porozdział 3.3), dopuszczam możliwość, że odrzucone wcześniej pojęcia reprezentacji (jako receptorów czy reprezentacji ukrytych) mogą zostać w przyszłości zmodyfikowane tak, by czynić zadość wymogowi opisu zadań.

Dlaczego pomimo tak znaczących zastrzeżeń decyduję się w ogóle na akceptację (I2)? Otóż nawet jeśli prowadzone tu rozważania nie wykazują, że teza (I2) jest prawdziwa ponad wszelką wątpliwość, to przynajmniej w zasadniczy sposób zmieniają one *sytuację dialektyczną*, w jakiej się znajdujemy. Ciężar argumentacji za tezą, że istnieją mechanizmy reprezentacyjne inne niż takie, które korzystają z wewnętrznych modeli, stoi po stronie ewentualnych jej zwolenników. Pod nieobecność takich alternatyw dla bronionej tu propozycji traktuję – choć, powtórzę, jedynie prowizorycznie – (I2) za twierdzenie z dużym prawdopodobieństwem prawdziwe.

h) Bycie MKM nie wystarcza do bycia mechanizmem reprezentacyjnym
Natura problemu: Czy spełnienie warunków nałożonych na bycie mechanizmem wyposażonych w konsumowany model rzeczywiście wystarcza do bycia mechanizmem reprezentacyjnym? Czy nie powinny być jeszcze spełnione inne, dodatkowe warunki, o których broniona tu koncepcja nie wspomina? Rzecz jasna sugestia ta naturalnie rodzi pytanie o potencjalnych „kandydatów" na taką dodatkową własność czy własności. Spróbujmy omawianemu zarzutowi nadać zatem bardziej konkretną postać. Jedna z możliwych własności reprezentacji, na jakie może powołać się krytyk, to zdolność do rozpoznawania błędów. Zdaniem niektórych autorów system korzystający z mechanizmów reprezentacyjnych musi mieć zdolność do rozpoznawania sytuacji, w których posługuje się on reprezentacją *błędną* (por.: Bickhard 2004a, 2004b; Anderson, Rosenberg 2008; Miłkowski 2013: 154–155). Podkreślę: nie chodzi tu o możliwość popełnienia błędu reprezentacyjnego, ale o możliwość rozpoznania, że taki błąd został popełniony. Wyjściowe zastrzeżenie można by za-

badają i opisują jego działanie za pomocą kategorii reprezentacyjnych. Wydaje się, że mechanizmy oparte na IO-reprezentacjach nie spełniają tego warunku. Innymi słowy, mamy tu do czynienia z pojęciem reprezentacji posiadającym zastosowanie instrumentalne/heurystyczne, lecz nie *eksplanacyjne*.

tem sformułować następująco: czy MKM nieposiadający możliwości rozpoznania błędnej reprezentacji (faktu posługiwania się błędną reprezentacją) zasługuje w ogóle na miano mechanizmu reprezentacyjnego?

Odpowiedź: Stoję na stanowisku, że spełnienie omówionych wcześniej warunków nakładanych na bycie MKM wystarcza do tego, aby dany mechanizm miał charakter reprezentacyjny. Dzieje się tak dlatego, że modele wykorzystywane w ramach tego rodzaju mechanizmów czynią zadość wymogowi opisu zadań. Spełnienie innych warunków nie wydaje się więc *konieczne*, aby uzsykać status mechanizmu reprezentacyjnego. Nie znaczy to jednak, że wyrażone wyżej sugestie są bezwartościowe.

Rozważmy możliwość, że „bycie systemem (mechanizmem) reprezentacyjnym" jest *stopniowalne*. Clark i Grush (1999) sugerują dla przykładu, że im bardziej oderwana reprezentacja, w tym większym stopniu reprezentacyjny będzie posługujący się nią system (tym „bardziej reprezentacyjny" będzie ten system). Idea, że mechanizmy czy systemy mogą być reprezentacyjne w mniejszym i większym stopniu, ma pewne zastosowanie także w kontekście naszych rozważań. Koncepcja MKM (w wersji, w jakiej została wyrażona w sekcji 4.2.1) może bowiem zostać potraktowana jako podająca warunki wystarczające do bycia *minimalnie* reprezentacyjnym mechanizmem. Dopuszczam jednak możliwość, że „czyste" MKM mogą mieć też inne własności, których posiadanie dodatkowo wzmacnia ich status jako mechanizmów reprezentacyjnych.

Można zilustrować powyższą ideę za pomocą wspomnianego wcześniej przykładu, czyli zdolności do rozpoznania błędu reprezentacyjnego. Musimy odróżnić sytuacje, w których teoria reprezentacji nie wspomina o kwestii detekcji błędu reprezentacyjnego, od sytuacji, w której teoria nie pozostawia (konceptualnie) możliwości dla zaistnienia takiej detekcji. Otóż do tej pory nie wymieniałem zdolności do rozpoznania błędu jako własności MKM. Sądzę jednak, że MKM mogą być wyposażone w zdolność do rozpoznania („rozpoznania") faktu, iż posługują się błędną reprezentacją. Zobaczmy, w jaki sposób okazuje się to możliwe.

Autorzy podkreślający wagę zdolności do detekcji błędów reprezentacyjnych twierdzą, że posiadanie tej zdolności jest ściśle powiązane z działaniowym czy pragmatycznym wymiarem reprezentacji (Bickhard 2004a, 2004b; Anderson, Rosenberg 2008)[40]. Rozwińmy tę ideę w kontekście interakcyjnej teorii reprezentacji Bickharda (2004a; 2004b). W ujęciu tego autora reprezentacje pozwalają na antycypację przebiegu potencjalnych działań systemu, a dokładniej – na określenie, które potencjalne działania czy interakcje ze środowiskiem zakończą się sukcesem. Błędną reprezentację poznaje się po tym, że działanie (reprezentowane jako) mające odnieść sukces zakończyło się niepowodzeniem. System może więc rozpoznać błąd reprezentacji wtedy, gdy podjęte na jej podstawie działanie (interakcja) nie zakończyło się sukcesem. Może to prowadzić do odrzucenia lub korekty tej reprezentacji. Czy tego rodzaju proces jest możliwy w kontekście MKM? Wydaje się, że tak.

Przypomnijmy sobie zaawansowany samochód Cumminsa, który wyposażono w pomniejszony model samego siebie i toru. Możemy sobie wyobrazić, że samochód ten dokonuje (na podstawie wewnętrznej symulacji) preselekcji mapy, którą wykorzysta, aby ukierunkować swój przejazd torem. Kiedy jednak dochodzi do właściwego przejazdu, okazuje się, że samochód zamiast posuwać się płynnie naprzód – uderza w krawędź toru. Brak sukcesu interakcyjnego może być wykorzystany jako wskazówka, że wybrana mapa nie odzwierciedla (w wystarczającym stopniu) struktury przestrzennej toru. Wyobraźmy sobie, że kiedy dochodzi do takiej nieudanej interakcji, samochód potrafi dokonać *korekty* antycypowanego przebie-

[40] Zdaniem Bickharda (2004a; 2004b) posiadanie takiej zdolności jest niemożliwe na gruncie teorii reprezentacji opartych na kodowaniu. Wedle tego autora z perspektywy koncepcji opartych na kodowaniu detekcja błędu wymagałaby sprawdzenia, czy między reprezentacją a tym, co reprezentowane, zachodzi odpowiedni rodzaj korespondencji. Jest to jednak zadanie niewykonalne, bo wymagające stanięcia niejako poza reprezentacją i porównania jej „z zewnątrz" do tego, co reprezentowane. Postulowanie takiej procedury jest obarczone regresem (potencjalnie *ad inifinitum*), ponieważ jej przeprowadzenie samo wymagałoby posługiwania się reprezentacjami (jak widać, Bickhard wykorzystuje tu pewien wariant jednego z klasycznych argumentów kierowanych przeciwko korespondencyjnej teorii prawdy).

gu zdarzeń. Na przykład wybiera on nową mapę spośród tych, które zostały wcześniej odrzucone lub przeprowadza symulację od nowa. Innymi słowy, nasz samochód potrafi rozpoznać i skorygować popełniony błąd reprezentacyjny, i to w sposób jak najbardziej zbieżny z pomysłami teoretycznymi Bickharda.

Nie wydaje się sprzeczne czy choćby kontrintuicyjne twierdzenie, iż reprezentacjami może posługiwać się mechanizm niezdolny do detekcji błędu. Na przykład nie ma nic sprzecznego czy kontrintuicyjnego w tezie, że z mechanizmu reprezentacyjnego korzysta organizm czy system sztuczny, który przy popełnieniu pierwszego błędu reprezentacyjnego natychmiast ginie (a zatem nie ma nigdy możliwości rozpoznania, że taki błąd popełnił). Właśnie z tego powodu twierdzę, iż nawet MKM niepozwalające w praktyce na detekcję błędu powinny być kwalifikowane jako mechanizmy reprezentacyjne. Jak jednak zobaczyliśmy, koncepcja MKM może zostać uzupełniona w taki sposób, aby dopuszczała możliwość rozpoznania błędu reprezentacyjnego. Przyznaję też, że korzystający z modelu mechanizm, który „potrafi" rozpoznać błąd, może zostać uznany za „mocniej" reprezentacyjny niż mechanizm, który takiej własności nie posiada. Twierdzę jednak, że przynajmniej do bycia mechanizmem *minimalnie* reprezentacyjnym wystarcza bycie MKM w wersji „czystej" czy „podstawowej" (opisanej w sekcji 4.2.1).

i) Dowolny MKM można opisać bez wykorzystywania kategorii reprezentacyjnych

Natura problemu: Teoretycznie jest możliwe opisanie dowolnego mechanizmu wyposażonego w konsumowany model na poziomie molekularnym, atomowym czy subatomowym. Jeśli tak, to dowolny MKM można opisać, nie odwołując się do modeli, nośników reprezentacji, ich konsumentów, treści intencjonalnej i tak dalej. Inaczej mówiąc, dowolny MKM możemy opisać w sposób, który całkowicie obywa się bez kategorii reprezentacyjnych. Nazwijmy zbiorczo takie alternatywne opisy – niereprezentacyjnymi. Na jakiej podstawie należy zatem uznawać za poprawny (czy preferowany) opis reprezentacyjny, a nie któryś z możliwych niereprezentacyjnych opisów MKM?

Odpowiedź: Odróżnijmy za Ramseyem (2007: 33–34)[41] trzy odrębne pytania:

> **1.** Czy jest możliwe opisanie działania mechanizmu *M* w kategoriach reprezentacyjnych?
> **2.** Czy jest absolutnie niezbędne opisanie działania mechanizmu *M* w kategoriach reprezentacyjnych?
> **3.** Czy istnieje jakiś eksplanacyjny zysk z opisywania działania mechanizmu *M* w kategoriach reprezentacyjnych?

Autor ten zauważa, iż odpowiedź na pytanie 1 jest zawsze (trywialnie) twierdząca. Jeśli nam na tym zależy, możemy posługiwać się terminologią reprezentacyjną na tyle swobodnie, by opisać za jej pomocą w zasadzie dowolny obiekt. Na przykład możemy bardzo liberalnie zastosować Dennettowską strategię intencjonalną (por. Dennett 2003) i przypisać leżącemu na drodze kamieniowi pragnienie pozostania w spoczynku, czyli pewien stan intencjonalny. W przypadku pytania 2 odpowiedź zawsze będzie z kolei (trywialnie) negatywna (Ramsey 2007: 33–34). *Prima facie dowolny* mechanizm możemy opisać niereprezentacyjnie, na przykład jako chmurę atomów. Terminologia reprezentacyjna nigdy nie jest zatem absolutnie *niezbędna* deskryptywnie.

Poszukując jednak koncepcji wyjaśniania reprezentacyjnego, nie zajmują mnie pytania o możliwość lub absolutną konieczność opisania mechanizmu w odpowiednich kategoriach. Kiedy mowa o mechanizmach reprezentacyjnych, chodzi o takie, w przypadku których to odpowiedź na *pytanie 3* jest twierdząca (por. Ramsey 2007). Są to mechanizmy, których opisanie w kategoriach reprezentacyjnych niesie ze sobą określone zyski eksplanacyjne; to znaczy, że dopiero opisując działanie tych mechanizmów w takich kategoriach, możemy wyjaśnić określone zjawisko. MKM są właśnie takim rodzajem mechanizmów. Nie zrozumiemy, jak MKM umożliwiają określone zjawiska poznawcze, o ile nie rozpoznamy i nie opiszemy faktu, że

[41] Wymienione pytania są sformułowane w sposób nieznacznie różniący się od tego, jak sformułował je Ramsey.

poszczególne ich komponenty realizują działania polegające na reprezentowaniu i konsumowaniu reprezentacji.

Aby lepiej zrozumieć tę odpowiedź na powyższy zarzut, warto odróżnić sytuacje, w których wiemy, *że* określony mechanizm wyjaśnia dane zjawisko, od sytuacji, w których wiemy, *jak* określony mechanizm wyjaśnia to zjawisko. Rozpatrzmy układ krwionośny jako mechanizm wyjaśniający dystrybucję składników odżywczych do tkanek organizmu. Możemy zapytać: dlaczego traktujemy układ krwionośny jako oparty na działaniu pompy (oraz innych funkcjonalnie scharakteryzowanych komponentów), zamiast zastosować opis fizykalny, który w ogóle nie odwołuje się do kategorii „pompowania"? Otóż dysponując opisem tego ostatniego rodzaju, dowiemy się być może, że złożony układ o określonych własnościach fizycznych umożliwia zachodzenie pewnego zjawiska (transportu składników odżywczych do tkanek organizmu). Taki opis nie powie nam jednak, jak ten układ fizyczny umożliwia to zjawisko. Wyjaśnienia mechanistyczne mają tymczasem dawać odpowiedź właśnie na to ostatnie pytanie. Aby taką odpowiedź uzyskać, musimy komponentom mechanizmu przypisać określone operacje, czyli funkcje. Dopiero rozkładając układ krwionośny na *funkcjonalnie* scharakteryzowane komponenty, zrozumiemy, jak układ krwionośny rozprowadza krew po organizmie – mianowicie wykorzystując (między innymi) swego rodzaju pompę (por. Craver 2001).

Powyższy sposób myślenia można zaaplikować do kwestii mechanizmów reprezentacyjnych. Jeśli chcemy wiedzieć, jak mechanizm wykorzystujący wewnętrzny model umożliwia pewne zjawisko, niewiele da nam dostarczenie opisu niereprezentacyjnego. Wyjaśnienie mechanistyczne (w przeciwieństwie do samego opisu) uzyskamy dopiero po przypisaniu komponentom mechanizmu *funkcji* nośnika i konsumenta reprezentacji. Jak to ujmują Clark i Grush, „opis w kategoriach reprezentacyjnych [...] stanowi klej, za pomocą którego do mechanizmu dołączamy *telos*" (1999: 8). Dopie-

ro opisane w kategoriach reprezentacyjnych – MKM pozwalają na wyjaśnianie określonych zjawisk[42].

4.3. Mechanizmy reprezentacyjne: zastosowania w kognitywistyce

Prowadzony dotychczas w tym rozdziale wywód był całkowicie podporządkowany zrealizowaniu głównego celu mojej pracy, czyli rozwiązaniu metaprzedmiotowego problemu statusu eksplanacyjnego reprezentacji mentalnych w kognitywistyce. Podejmując jednak kwestię tego, na czym polega mechanistyczne wyjaśnienie reprezentacyjne oraz skupiając się na uproszczonych, stworzonych na potrzeby argumentacji przykładach tego rodzaju wyjaśnień, całkowicie pominąłem tu zagadnienie, czy system poznawczy *rzeczywiście* jest systemem reprezentacyjnym. To ostatnie zagadnienie dotyka zaś sedna problemu, który w rozdziale 1 został przeze mnie określony jako *przedmiotowy* problem statusu eksplanacyjnego reprezentacji w naukach kognitywnych.

Moim celem w tym podrozdziale będzie prześledzenie konsekwencji, jakie proponowane tu rozwiązanie metaprzedmiotowego problemu reprezentacji niesie ze sobą dla problemu przedmiotowego. Wcześniej, w rozdziale 1 (podrozdział 1.3) zostało przecież zaznaczone, iż zasadniczym teoretycznym „zyskiem" z rozwiązania problemu metaprzedmiotowego powinno być właśnie to, że uzyskamy w ten sposób dokonania postępu przy rozwiązywaniu problemu przedmiotowego. A zatem wiedząc, na czym polega wyjaśnianie reprezentacyjne w naukach kognitywnych, powinniśmy być zdolni do udzielenia – w sposób unikający arbitralności, uznaniowości i niejasnych presupozycji – odpowiedzi na pytanie o to, czy (lub w jakim

[42] Pamiętajmy, że podstawę tego zysku eksplanacyjnego stanowi realna, niezależna od ludzkich praktyk epistemicznych przyczynowo-funkcjonalna organizacja mechanizmu. Na przykład *opisanie* komponentu mechanizmu jako konsumenta pozwala *wyjaśnić* dane zjawisko tylko wtedy, gdy komponent ten rzeczywiście (niezależnie od obserwatora) „zajmuje" się w mechanizmie konsumowaniem reprezentacji.

zakresie) system poznawczy jest systemem reprezentacyjnym. Chcę teraz wykorzystać teorię mechanizmów wyposażonych w konsumowane modele, aby zarysować taką odpowiedź.

W kontekście dotychczasowych ustaleń – pytanie o to, czy system poznawczy jest systemem reprezentacyjnym, powinno zostać uszczegółowione jako pytanie o to, czy działanie tego systemu opiera się na MKM. Na tym polega, jak sądzę, zasadnicza wartość dotychczasowych rozstrzygnięć poczynionych w tej książce. Pozwalają nam one doprecyzować i rozjaśnić, o co pytamy, kiedy zastanawiamy się nad rzeczywistą rolą reprezentacji w systemie poznawczym. Postawić jednak tezę, że system poznawczy jest reprezentacyjny o ile wykorzystuje MKM, to jedno, a stwierdzić, że *de facto* korzysta on z takich mechanizmów – to co innego. Czy proponowana tu teoria mechanizmów reprezentacyjnych opisuje coś więcej niż jedynie „zabawkowe" przykłady w rodzaju samochodu Cumminsa lub Bezmyślnego Jana? Czy stan teoretyczny współczesnej kognitywistyki daje podstawy, by sądzić, iż realne systemy poznawcze działają na podstawie mechanizmów korzystających z wewnętrznych, konsumowanych modeli?

Dyskusję powyższego zagadnienia warto rozpocząć od omówienia diagnozy przedstawionej przez wspominanego tu wielokrotnie Ramseya (2007). Jak pamiętamy, jego *Representation Reconsidered* to praca o wyraźnie antyreprezentacjonistycznym wydźwięku. Przedstawione tam rozważania prowadzą autora do postawienia tezy, że najlepsze dostępne kognitywistyczne teorie działania systemu poznawczego obywają się bez postulowania (rzeczywistych, a nie jedynie nominalnych) struktur pełniących funkcje reprezentacji. Przyjrzymy się temu, na jakiej podstawie Ramsey wyciąga taką konkluzję.

Zasadniczą rolę w wywodzie Ramseya (2007: 2–4, 189–235) odgrywa rozróżnienie między kognitywistyką „klasyczną" a „nieklasyczną". Ta pierwsza kategoria obejmuje tak zwaną Starą Dobrą Sztuczną Inteligencją (GOFAI, *Good Old-Fashioned Artifical Intelligence*), czyli podejście w naukach kognitywnych oparte na rozumieniu systemu poznawczego jako „klasycznie" obliczeniowego, to znaczy przetwarzającego symbole zgodnie z formalnymi, syntaktycznymi regułami (na ogół *explicite* kodowanymi w systemie). Do nieklasycznej kogni-

tywistyki zalicza zaś Ramsey wszystkie podejścia do badania, modelowania i wyjaśniania aktywności systemu poznawczego, jakie wyłoniły się po okresie dominacji GOFAI i które pod jakimiś istotnymi względami odchodzą od idei, że system poznawczy ma naturę symboliczno-obliczeniową. Do grupy tej moglibyśmy zaliczyć takie podejścia w kognitywistyce, jak: koneksjonizm, podejście wyznaczane przez neuronaukę poznawczą i obliczeniową, podejście oparte na teorii systemów dynamicznych, podejście oparte na teorii sterowania czy niektóre programy z zakresu robotyki poznawczej[43]. Bardzo istotne znaczenie dla wywodu Ramseya (2007: 189–235) zawarte w *Representation Reconsidered* twierdzenie, że definitywnie skończyły się czasy, w których GOFAI wyznaczało główny nurt teoretyczno-badawczy kognitywistyki i w których można było mieć nadzieję, że właśnie na gruncie tego podejścia wyrośnie kompletna, zadowalająca teoria działania systemu poznawczego. W ujęciu Ramseya klasyczna kognitywistyka jest obecnie skutecznie wypierana przez nowsze podejścia, a proces ten rozpoczął się wraz z powstaniem koneksjonizmu. GOFAI zostało lub już niedługo zostanie jednoznacznie zastąpione przez jakieś podejście nieklasyczne. Innymi słowy, perspektywa symboliczno-obliczeniowa należy już tylko do historii kognitywistyki.

Zawężę teraz rekonstrukcję argumentacji Ramseya w taki sposób, aby skoncentrować się tylko na interesującym mnie rodzaju reprezentacji, czyli modelach mentalnych (wewnętrznych S-reprezentacjach). Autor ten przyznaje, że pojęcie reprezentacji jako modeli mentalnych odgrywało ważną rolę w ramach GOFAI. Przytacza on szereg klasycznych koncepcji funkcjonowania systemu poznawczego – w ich świetle system ten miał korzystać z mechanizmów obliczeniowych, które moglibyśmy sklasyfikować jako MKM. Ramsey

[43] Lista ta nie pretenduje do miana rozłącznej klasyfikacji. Wiele konkretnych teorii, modeli czy wyjaśnień sformułowanych przez kognitywistów można by bez wątpienia zaliczyć do kilku z wymienionych podejść. Także same wymienione podejścia są często ściśle ze sobą związane. Na przykład podejście odwołujące się do teorii sterowania wykorzystuje aparaturę pojęciową teorii systemów dynamicznych, a neuronauka obliczeniowa i robotyka poznawcza korzystają z osiągnięć koneksjonizmu.

wymienia tu między innymi wspomnianą już wcześniej teorię modeli mentalnych Philipa Johnsona-Lairda, a także takie symboliczno-obliczeniowe koncepcje funkcji poznawczych, jak model SOAR Johna Lairda, Allena Newella i Paula Rosenblooma (1987) czy ACT Johna Andersona (1996). Jednak zdaniem tego autora (Ramsey 2007: 79–80) znaczenie pojęcia reprezentacji jako wewnętrznych modeli w kognitywistyce zaczyna i kończy się na GOFAI. W ramach nieklasycznej kognitywistyki takiego rodzaju reprezentacje nie są już w ogóle postulowane[44]. Tym samym otrzymujemy werdykt antyreprezentacjonistyczny, ponieważ okazuje się, iż prawomocne pojęcie reprezentacji (jako modeli mentalnych) należy już tylko do historii nauk kognitywnych. Rozumowanie Ramseya jest zatem następujące:

1. Pojęcie reprezentacji jako modeli (S-reprezentacji) spełnia wymóg opisu zadań, czyli odwołuje się ono do eksplanacyjnie prawomocnego czy wartościowego rodzaju reprezentacji mentalnych.
2. Pojęcie reprezentacji jako modeli było używane w ramach klasycznej kognitywistyki, jednak nie odgrywa żadnej roli w nieklasycznej kognitywistyce (nieklasyczne teorie nie postulują istnienia modeli mentalnych).
3. Klasyczna kognitywistyka to zdezaktualizowane podejście do wyjaśniania działania systemu poznawczego; została ona (lub wkrótce zostanie) całkowicie zastąpiona przez wersję nieklasyczną.
4. Zatem pojęcie reprezentacji jako modeli jest teoretycznie zdezaktualizowane – nie odgrywa/nie będzie odgrywać ono żadnej roli w poprawnej teorii systemu poznawczego, sformułowanej na gruncie jednego z podejść nieklasycznych.

[44] Ramsey wspomina niemal mimochodem (2007: 80, przypis 5), że istnieje w nieklasycznej kognitywistyce „kilka" (*a few*) teorii wykorzystujących pojęcie reprezentacji jako modeli. Jednak najwyraźniej uznaje on je za tak rzadkie i peryferyjne, że ich istnienie nie unieważnia stawianej przez niego diagnozy dotyczącej obecnego stanu teoretycznego kognitywistyki. Sądzę, że taka ocena sytuacji jest wyrazem „niedoszacowania" przez Ramseya rzeczywistej roli pełnionej współcześnie przez pojęcie reprezentacji mentalnych jako modeli w naukach kognitywnych.

5. Zatem współczesna (nieklasyczna) kognitywistyka przedstawia system poznawczy jako niekorzystający z wewnętrznych reprezentacji[45].

Chcę odrzucić antyreprezentacjonistyczną konstatację Ramseya. Chociaż zgadzam się z niektórymi rozwiązaniami bronionymi przez tego autora na poziomie metaprzedmiotowym – konkretnie podzielam jego werdykt w sprawie modeli mentalnych jako eksplanacyjnie wartościowego rodzaju reprezentacji – to uznaję jego antyreprezentacjonizm na poziomie przedmiotowym za nieuzasadniony. Uważam, że utożsamienie reprezentacji w kognitywistyce z wewnętrznymi modelami zdecydowanie sprzyja reprezentacjonizmowi. Chcę bronić reprezentacjonizmu na poziomie przedmiotowym, pokazując, że argumentacja Ramseya za stanowiskiem przeciwnym jest zupełnie nieskuteczna.

Przedstawiony wyżej argument można próbować „rozmontować" na (co najmniej) dwa sposoby. Po pierwsze, można by twierdzić, że Ramseyowskie „pożegnanie" GOFAI jest przedwczesne (zanegować przesłankę 3). Po drugie, można by twierdzić, że fałszywa jest teza Ramseya, iż pojęcie S-reprezentacji nie odgrywa ważnej roli w nieklasycznej kognitywistyce (zanegować przesłankę 2). Moim celem w dalszej części tego podrozdziału będzie zrealizowanie tej ostatniej strategii. Nie chcę tu polemizować z twierdzeniem o postępującej marginalizacji GOFAI (choć nie wykluczam, że można z nim polemizować), a zamiast tego skupię się na pokazaniu, że to, co Ramsey uznaje za nieklasyczną kognitywistykę, jest pełne wyja-

[45] Formułując wniosek 5, zakładam, że modele mentalne są jedynym eksplanacyjnie prawomocnym rodzajem reprezentacji. Bez przyjęcia takiego założenia konkluzja 5 nie wynikałaby z 4, ponieważ mogłoby się okazać, że nieklasyczna kognitywistyka postuluje inne niż modele rodzaje prawomocnych eksplanacyjnie reprezentacji. Warto mieć na uwadze, że sam Ramsey (2007: 68–77) dopuszcza jeszcze IO-reprezentacje jako eksplanacyjnie uprawnione (twierdzenie to sam odrzucam – por. przypis 39). Należy jednak zaznaczyć, że także ten rodzaj reprezentacji uznaje on za ściśle powiązany z GOFAI, a przez to mający wartość jedynie historyczną (Ramsey 2007). Innymi słowy, w ujęciu Ramseya argument strukturalnie identyczny z wymienionym powyżej można zastosować do IO-reprezentacji.

śnień odwołujących się do wewnętrznych modeli (por.: Grush 2008; Calvo, García 2009; Sprevak 2011). Wbrew temu przekonaniu pojęcie reprezentacji jako modeli mentalnych jest jak najbardziej obecne i stosunkowo rozpowszechnione w nieklasycznej kognitywistyce. Twierdzenie, że współczesna kognitywistyka obywa się bez takiego narzędzia eksplanacyjnego jak reprezentacje mentalne, jest zatem po prostu fałszywe i bezzasadne.

Poniższa lista zawiera kontrprzykłady dla stawianej przez Ramseya diagnozy współczesnego stanu nauk kognitywnych. Pokazuje ona przypadki wykorzystania pojęcia modeli wewnętrznych w ramach różnych nieklasycznych podejść teoretyczno-badawczych w kognitywistyce. Nie twierdzę, że wymienione przykłady zawsze już na pierwszy rzut oka idealnie wpasowują się w przyjmowany tu sposób rozumienia mechanizmów reprezentacyjnych (jako MKM). Za każdym razem postaram się zatem krótko rozjaśnić, na jakiej podstawie uważam, że w danym przypadku wykorzystuje się pojęcie konsumowanego modelu (czy też postuluje się istnienie takiego modelu).

a) MKM w koneksjonizmie: Cottrella sieć dyskryminująca twarze

Według Ramseya (2007: 226–229) trend odchodzenia od reprezentacjonizmu w kognitywistyce rozpoczął się wraz z powstaniem i wzrostem znaczenia koneksjonizmu. Choć koneksjoniści nie przestali posługiwać się terminem „reprezentacja", to przestali w istocie powoływać się na struktury rzeczywiście pełniące funkcję polegającą na reprezentowaniu czegoś. Koneksjonizm operuje bowiem według Ramseya nieuprawnionymi eksplanacyjnie – bo nieprzechodzącymi testu opartego na wymogu opisu zadań – pojęciami reprezentacji jako receptorów oraz reprezentacji ukrytych. Przy bliższym, skrupulatnym spojrzeniu koneksjonizm okazuje się zatem według tego autora całkowicie antyreprezentacjonistycznym podejściem do modelowania systemu poznawczego.

Jak się jednak wydaje, Ramsey nie do końca trafnie odczytuje rolę, jaką odgrywa pojęcie reprezentacji w koneksjonizmie (O'Brien, Opie 2004; Calvo, García 2009; Shagrir 2012). Okazuje się bowiem, że przynajmniej niektóre sieci koneksyjne działają na podstawie we-

wnętrznych struktur grających rolę nie tyle receptorów albo reprezentacji ukrytych, ile raczej wewnętrznych *modeli*. Jako ilustrację tego stanu rzeczy można wskazać stworzoną przez Garrisona Cottrella sieć „rozpoznającą" (czyli selektywnie reagującą na) ludzkie twarze przedstawione na fotografiach (Churchland 2002; O'Brien, Opie 2004).

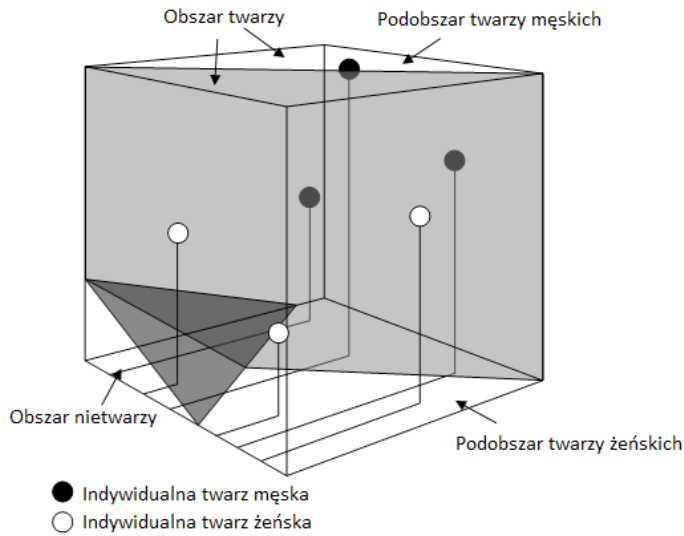
W okresie treningowym sieć Cottrella była eksponowana na 64 monochromatyczne zdjęcia 11 ludzkich twarzy (podpisanych za każdym razem imieniem przedstawionej na fotografii osoby) oraz 13 zdjęć nieprzedstawiających twarzy (tu i dalej: Churchland 2002; O'Brien, Opie 2004). Była ona trenowana w kierunku nabycia zdolności do (1) odróżniania twarzy od nietwarzy; (2) odróżniania płci osoby przedstawionej na zdjęciu; (3) „rozpoznawania" tożsamości (imienia) osoby przedstawionej na zdjęciu. Pierwszą, wejściową warstwę sieci stanowiła sztuczna „siatkówka", czyli układ złożony z 64 × 64 (razem 4096) komórek. Każdy z neuronów w tej warstwie odpowiadał jednemu pikselowi zdjęcia i mógł przyjmować jeden z 256 poziomów aktywności (odpowiadających jasności danego piksela). Druga, pośrednicząca warstwa miała 80 komórek i była połączona z obejmującą osiem komórek warstwą wyjściową, generującą „rozstrzygnięcia" sieci w sprawie (1) statusu obiektu przedstawionego na zdjęciu jako twarzy lub nietwarzy; (2) płci przedstawionej na zdjęciu osoby; (3) tożsamości (imienia) przedstawionej osoby. Jak się okazało, po pewnym czasie sieć osiągnęła 100% skuteczności dyskryminacyjnej przy eksponowaniu jej na zdjęcia z sesji treningowej, 98% skuteczności przy rozpoznawaniu zupełnie nowych zdjęć tych samych osób, a także wysoką skuteczność (odpowiednio 100% i 80%) przy odróżnianiu twarzy od nietwarzy oraz rozpoznawaniu płci na fotografiach przedstawiających zupełnie nowe osoby.

Nietrudno zrozumieć, w jaki sposób działanie sieci Cottrella mogłoby zostać zinterpretowane przez pryzmat pojęcia reprezentacji jako receptorów (wzorce aktywności neuronów w sieci systematycznie współzmieniają się z określonymi własnościami zdjęć) czy pojęcia reprezentacji ukrytych (struktura wag połączeń synaptycznych sieci miałaby „ucieleśniać" ukryte, rozproszone i nałożone reprezentacje twarzy). Biorąc pod uwagę rozstrzygnięcia poczynione

w rozdziale 3 (podrozdział 3.3), możliwość interpretacji działania sieci w taki sposób wcale nie zapewnia jej statusu systemu czy mechanizmu reprezentacyjnego. Jak jednak zauważają Gerard O'Brien i Jon Opie (2004), istnieje jeszcze jeden rodzaj reprezentacji, do którego można się odwołać, wyjaśniając sukces detekcyjny sieci Cottrella. Chodzi o S-reprezentację, czyli reprezentację jako (wewnętrzny) model. Kluczowe znaczenie ma tu zwrócenie uwagi na funkcję, jaką pełni w wytrenowanej sieci Cottrella warstwa pośrednicząca, czyli ta złożona z 80 sztucznych neuronów. Kiedy przedstawić aktywność tej warstwy jako trajektorię w wielowymiarowej przestrzeni aktywacji (przestrzeni stanów) – przestrzeni, której poszczególne wymiary są wyznaczane przez możliwe poziomy aktywności każdego z poszczególnych neuronów tej warstwy (por. uproszczona, trójwymiarowa ilustracja tej przestrzeni na rysunku 6) – to okaże się, że zachodzi *strukturalne podobieństwo* między tą przestrzenią aktywacji (jej własnościami metrycznymi) a strukturą fizycznych podobieństw zachodzących między twarzami, na które sieć jest eksponowana. Każdemu punktowi w takiej przestrzeni odpowiada zatem jedna twarz, a struktura przestrzennych odległości między punktami tej przestrzeni odzwierciedla strukturę podobieństw fizycznych zachodzących między twarzami. Punkty tej przestrzeni znajdują się tym bliżej siebie, im bardziej podobne są odpowiadające im twarze. Co więcej, przestrzeń aktywacji drugiej warstwy dzieli się na obszary i podobszary odpowiadające poszczególnym kategoriom (twarze/nietwarze) i subkategoriom (twarze męskie/żeńskie) bodźców (por. rysunek 6). Zachodzenie takiego podobieństwa drugiego rzędu *prima facie* sprzyja uznaniu warstwy pośredniczącej za nośnik S-reprezentacji, której przedmiotem są twarze, na jakie jest eksponowana sieć (struktura podobieństw między twarzami).

Oczywiście samo zachodzenie podobieństwa strukturalnego nie czyni jeszcze z czegoś modelu. Model musi być *wykorzystany* jako model w szerszym mechanizmie. Sieć Cottrela wydaje się jednak spełniać i ten warunek. Jak zauważają O'Brien i Opie (2004), przestrzeń aktywacji środkowej warstwy sieci nie jest jedynie epifenomenalnie podobna do przestrzeni podobieństw między twarzami. Wydaje się, że to właśnie zachodzenie owego podobieństwa zapewnia

sieci jej zdolność do odpowiedniej, selektywnej reakcji na fotografie. Inaczej mówiąc, sieć osiąga założone przez projektanta cele dzięki posiadaniu – w warstwie pośredniczącej – S-reprezentacji pewnej domeny (stanowiącej przedmiot reprezentacji).



Rysunek 6. Podziały w przestrzeni aktywacji drugiej warstwy neuronów w sieci koneksyjnej Cottrella. Źródło: Churchland 2002: 61 oraz O'Brien, Opie 2004: 16

Powyższe twierdzenie wymaga pewnego uzupełnienia. Mamy bowiem do czynienia z następującą sytuacją: opisywana sieć wydaje się intuicyjnie korzystać z modelu, jednak trudno w niej zlokalizować komponent, który moglibyśmy utożsamić z konsumentem tego modelu. Przypisanie roli konsumenta wyjściowej (trzeciej) warstwie sieci może zostać bowiem uznane za zbyt liberalne i wykonane *ad hoc*. Warstwa wyjściowa nie „robi" w sieci nic poza wykazywaniem aktywności w jakiś sposób skorelowanej z różnymi własnościami przedstawianych na wejściu zdjęć. Sądzę jednak, że problem „braku konsumenta" nie będzie tu nierozwiązywalny. Aby dosto-

sować omawianą sieć do teorii MKM, weźmy pod uwagę, że docelowo powinna ona stanowić część jakiegoś szerszego mechanizmu poznawczego. W takim mechanizmie aktywność ostatniej, wyjściowej warstwy sieci nie będzie funkcjonalnie jałowa (nie będzie współzmieniać się z bodźcami tylko po to, by zrealizować założenia konstruktora). Zamiast tego owa warstwa zostanie sprzężona z jakimś funkcjonalnie wyspecjalizowanym komponentem, na przykład komponentem motorycznym, którego zadanie polega na wygenerowaniu reakcji behawioralnej dostosowanej do tożsamości osoby rozpoznanej na zdjęciu[46]. Właśnie taki komponent pełniłby w sieci rolę konsumenta modelu. Sieć Cottrella w takiej sytuacji generowałaby wewnętrzny model twarzy (zlokalizowany czy zaimplementowany w drugiej, pośredniczącej warstwie neuronów), zaś komponent motoryczny korzystałby z podobieństwa między modelem a modelowaną domeną w celu wygenerowania stosownych do sytuacji zachowań.

b) MKM w teorii sterowania: teoria emulacji

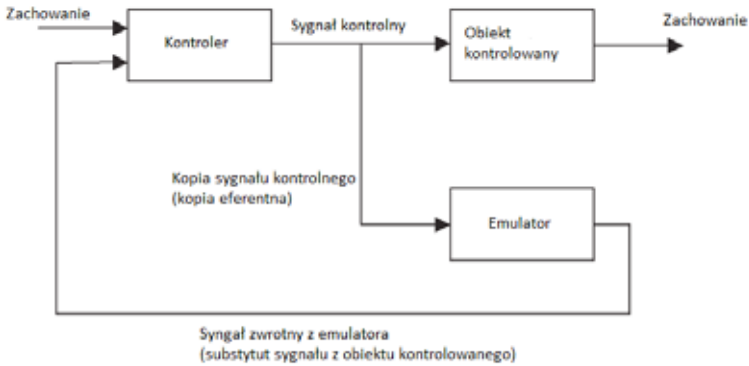
Jednym z najbardziej znaczących i klarownych przykładów obecności mechanizmów wykorzystujących wewnętrzne modele w nieklasycznej kognitywistyce jest teoria emulacji autorstwa Ricka Grusha (1997; 2004). Koncepcja ta wywodzi się z podejścia w naukach kognitywnych, które opiera się na teorii sterowania, czyli dyscyplinie z pogranicza matematyki i inżynierii, zajmującej się problemem kontroli w systemach dynamicznych. Teoria emulacji stanowi przede wszystkim właśnie propozycję dotyczącą sterowania. Ma ona wyjaśnić, jak system poznawczy radzi sobie z zadaniem polegającym na płynnym i odpowiednio szybkim kontrolowaniu zachowań motorycznych ciała własnego (precyzyjniej – systemu mięśniowo-szkieletowego).

Centralne znaczenie dla koncepcji Grusha (1997, 2004) ma obserwacja dotycząca tego, w jaki sposób problem kontroli motorycznej nie może zostać rozwiązany. Mianowicie proces płynnej

[46] Można wręcz powiedzieć, że dopiero w takiej sytuacji będziemy w ogóle mieli do czynienia z procesem zasługującym na miano „rozpoznania" czy „detekcji" twarzy, a nie ze zwykłym procesem selektywnej reakcji na bodźce.

i sprawnej kontroli ruchu nie może odbywać się na podstawie ani (1) mechanizmu opartego na sprzężeniu wyprzedzającym (czyli na wysłaniu kompletnej komendy motorycznej do efektorów), ani (2) mechanizmu opartego na sprzężeniu zwrotnym. Opcja (1) wymagałaby postulowania nierealistycznie precyzyjnego z punktu widzenia biologii systemu planowania ruchu. Opcja (2) jest wykluczona ze względu na opóźnienie między wygenerowaniem komendy motorycznej a dojściem do systemu motorycznego sygnału zwrotnego o przebiegu ruchu. Jak zatem problem kontroli ruchu zostaje rozwiązany w realnych, biologicznych systemach poznawczych? Sedno propozycji Grusha stanowi twierdzenie o istnieniu w ośrodkowym układzie nerwowym struktur nazywanych *emulatorami*. Mówiąc najprościej, emulatory to nic innego jak wewnętrzne, neuronalne modele systemu mięśniowo-szkieletowego. W ujęciu Grusha – kontrolując ruch, mózg wykorzystuje coś w rodzaju wirtualnej „makiety" systemu mięśniowo-szkieletowego, dzięki czemu może on z odpowiednim wyprzedzeniem antycypować skutki poszczególnych sekwencji ruchowych. W odniesieniu do tych predykcji jest możliwe adaptywne modyfikowanie wykonywanych ruchów, jeszcze zanim do systemu motorycznego dotrą sygnały zwrotne z ciała. Za każdym razem, gdy do efektorów zostanie wysłana komenda motoryczna, do emulatora (emulatorów) jest wysyłana jej „kopia eferentna" (por. rysunek 7). Aktywność emulatora pod wpływem kopii eferentnej ma docelowo imitować (dynamicznie odzwierciedlać, symulować) aktywność w ramach pętli system motoryczny–efektory i dzięki temu generować zastępczy sygnał zwrotny, obejmujący informacje proprioceptywne i kinestetyczne, jak również sygnały w innych modalnościach zmysłowych, w tym sygnał wzrokowy. Zamiast „czekać" na informacje zwrotne z ciała (i świata) ośrodkowy układ nerwowy wykorzystuje emulator jako komponent generujący na bieżąco „udawane" sygnały. Na podstawie takich predykcji są generowane kolejne komendy motoryczne. W ujęciu Grusha emulatory to zatem wewnętrzne modele umożliwiające mózgowi predykcję

przebiegu ruchu, natomiast kontrola motoryczna opiera się na realizowanym przez emulatory procesie „kontrolowanej halucynacji”[47].



Rysunek 7. Proces emulacji w ujęciu Ricka Grusha. Kontroler (komponent wysyłający komendy motoryczne do efektorów) steruje obiektem (ciałem), wykorzystując predykcje ustawicznie generowane (na podstawie kopii eferentnych komend motorycznych) przez emulator. Źródło: Grush 2004: 379

Grush (1997, 2004) odróżnia w swojej teorii dwa rodzaje emulatorów: (1) emulator amodalny, czyli zmysłowo niespecyficzny, zajmujący się emulowaniem (modelowaniem) organizmu i środowiska za pomocą czysto przestrzennych własności określonych w egocentrycznej ramie odniesienia, oraz (2) emulatory modalne, czyli powiązane z poszczególnymi modalnościami zmysłowymi, które mogą generować zmysłowo specyficzne „zastępcze” sygnały zwrot-

[47] Grush (1997, 2004) nie twierdzi, że właściwe (niebędące „halucynacjami”) sygnały zwrotne z ciała i środowiska nie odgrywają w kontroli motorycznej ważnej roli. Jest dokładnie odwrotnie – sygnały zwrotne pełnią w tym procesie zasadniczą funkcję. Są one na bieżąco rejestrowane, porównywane z wynikami emulacji i wykorzystywane do korekty ruchu, jeśli rzeczywisty jego przebieg odbiega od przebiegu, który został przewidziany za pomocą emulatora (por. wspomniana w tekście głównym teza o emulatorach jako filtrach Kalmana).

ne[48]. W jaki jednak sposób emulatorom – zarówno modalnym, jak i amodalnym – udaje się „naśladować" działanie układu mięśniowo-szkieletowego? Grush (2004) wskazuje dwie potencjalne odpowiedzi, które uznaje za dopuszczalne z punktu widzenia własnej teorii. Z jednej strony emulator może stanowić po prostu formę pamięci składującej czysto asocjacyjne związki między różnymi możliwymi komendami motorycznymi a możliwymi sensorycznymi sygnałami zwrotnymi (imitując tym samym korelacje wejście–wyjście w ramach pętli system motoryczny–efektory). Z drugiej strony emulator może być „artykułowany" (*articulated*), to znaczy złożony z komponentów określanych za pomocą zmiennych. Każda zmienna opisująca dany komponent emulatora miałaby odpowiadać konkretnej własności systemu mięśniowo-szkiletowego, na przykład możliwym kątom zginania określonego stawu. Wzajemne zależności między komponentami emulatora artykułowanego (opisywane przez określone prawa dynamiczne) miałyby dynamicznie odzwierciedlać zależności (także opisywane przez określone prawa dynamiczne) między elementami układu mięśniowo-szkieletowego.

Niezależnie od tego, czy emulatory są artykułowane, czy też stanowią formę pamięci asocjacyjnej, według Grusha (2004) zawsze pełnią one w systemie funkcję filtru Kalmana. Nie wdając się

[48] Grush spekuluje, że emulator amodalny wchodzi w interakcje z emulatorami modalnymi: „idea polega na tym, że organizm posiadający, powiedzmy, dwa emulatory modalne, wzrokowy i słuchowy, mógłby ich używać równocześnie z emulatorem amodalnym. W takim wypadku amodalny emulator mógłby podlegać dwóm rodzajom »pomiaru«: pomiarowi wzrokowemu, dającemu predykcję dotyczącą tego, co powinno zostać zobaczone, biorąc pod uwagę szacowany [przez emulator amodalny – P. G.] stan środowiska; oraz pomiarowi słuchowemu, dającemu predykcję dotyczącą tego, co powinno zostać usłyszane, biorąc pod uwagę szacowany stan środowiska. Emulator amodalny byłby uaktualniony przez oba te rezydua sensoryczne [czyli różnice między apriorycznie oszacowaną wartością sygnału a jego rzeczywistą wartością – P. G.], tworząc oszacowanie stanu [czyli aprioryczne oszacowanie wartości sygnału dokonane przy założeniu, że nie dochodzi do żadnego nieprzewidzianego szumu – P. G.], które integruje informacje z wszystkich modalności, jak również aprirorycznе oszacowanie stanu środowiska [...]" (2004: 389). Mówiąc prościej, oba rodzaje emulatorów współpracują ze sobą: emulatory modalne uzupełniają wyniki działania emulatora amodalnego.

w szczegóły matematyczne, idea polega na tym, że na podstawie aktywności emulatorów można określić różnicę między rzeczywistym sygnałem zwrotnym dochodzącym z ciała a sygnałem zwrotnym (stanowiącym właśnie efekt aktywności emulatorów), który powinien być oczekiwany w sytuacji niewystępowania szumów wynikających z okoliczności zewnętrznych i zakłóceń pomiarowych. Emulatory pozwalają zatem odróżnić wewnętrzną dynamikę procesu kontroli motorycznej od nieprzewidzianych zakłóceń o charakterze zewnętrznym lub pomiarowym.

Należy wyraźnie zaznaczyć, że wartość eksplanacyjna teorii emulacji nie kończy się na dostarczeniu (potencjalnego) eksplanandum zjawiska kontroli motorycznej. Według Grusha oraz innych autorów koncepcja emulacji daje także wgląd w naturę percepcji (Grush 2004; Piłat 2006), naturę wyobraźni percepcyjnej i przestrzennej (za pierwszą odpowiadałyby poszczególne emulatory modalne, a za drugą – emulator amodalny; Grush 2004), naturę zdolności czytania umysłów (Gardenförs 2007) oraz niektóre zjawiska związane z doświadczeniem ciała własnego (Grush 2004; Nowakowski 2010). Ze względów objętościowych pomijam tu kwestię danych empirycznych świadczących za istnieniem emulatorów. Dość jednak powiedzieć, że teoria emulacji (1) dobrze wpisuje się w wyniki eksperymentów pokazujących, iż procesy wyobraźniowe wiążą się z aktywnością *off-line* neuronalnych obszarów motorycznych i sensorycznych, oraz (2) jest zgodna z danymi empirycznymi dotyczącymi kończyn fantomowych (por. Grush 2004). Na gruncie teorii emulacji można uznać, że oba wymienione zjawiska – aktywność motoryczna i percepcyjna *off-line* oraz doświadczenie kończyn fantomowych – wynikają właśnie z aktywności emulatorów.

Choć uważam związek między teorią emulacji a koncepcją MKM za dość ewidentny, warto dokładnie zaznaczyć, na czym on polega. Postulowane przez Grusha emulatory stanowią bardzo klarowny przykład modeli wewnętrznych. Poprawne funkcjonowanie emulatora w systemie poznawczym jest ściśle zależne od tego, jak wiernie proces emulacji odzwierciedla (dynamicznie) aktywność układu mięśniowo-szkieletowego (czy też pętli system motoryczny–efektory). Emulator będzie tym bardziej funkcjonalny dla sys-

temu, im lepsze predykcje zapewnia; a zapewnia tym lepsze predykcje, im wierniej odzwierciedla (symuluje) działanie emulowanego obiektu czy procesu. Który komponent systemu poznawczego *konsumuje* jednak aktywność emulatorów? Kiedy skupimy się wyłącznie na roli emulatorów w kontroli motorycznej, to prosta analiza typu *Cui bono?* pokazuje, że za konsumenta powinniśmy uznać kontroler, czyli komponent systemu motorycznego wyspecjalizowany funkcjonalnie w generowaniu, a następnie wysyłaniu do ciała komend motorycznych. Zgodnie z teorią Grusha to właśnie kontroler wykorzystuje emulator jako *surogat* kontrolowanego systemu. Innymi słowy, emulator pozwala kontrolerowi „podejmować decyzje" motoryczne pod nieobecność sygnałów dochodzących bezpośrednio z przedmiotu reprezentacji, czyli układu mięśniowo-szkieletowego.

c) MKM w neuronauce poznawczej: teoria symboli percepcyjnych

Kolejnym przykładem wykorzystania MKM jako narzędzia eksplanacyjnego w nieklasycznej kognitywistyce jest rozwijana na gruncie neuronauki poznawczej teoria symboli percepcyjnych (dalej: TSP). TSP została sformułowana przez psychologa Lawrence'a Barsalou (1999) i rozwinięta z perspektywy filozofii umysłu przez filozofa Jessego Prinza (2002)[49]. TSP to koncepcja pojęć rozumianych jako mentalne reprezentacje kategorii, które uczestniczą w realizowaniu typowo pojęciowych funkcji poznawczych, takich jak kategoryzacja czy przeprowadzanie rozumowań. Centralna jej teza głosi, iż wykorzystywanie reprezentacji pojęciowych opiera się na działaniu zmysłowo specyficznych (modalnych) obszarów percepcyjnych oraz systemu motorycznego (Barsalou 1999; Prinz 2002). Dokładniej mówiąc, zgodnie z TSP (1) nośnikami pojęć są systemy neuronalne pierwotnie odpowiedzialne za funkcje percepcyjne oraz kontrolę motoryczną, a także – (2) operowanie pojęciami wiąże się z (re)

[49] Należy zaznaczyć, że TSP w odmianie zaproponowanej przez Prinza odbiega pod jednym względem od bronionej tu koncepcji MKM. Otóż w jego ujęciu (Prinz 2002: 237–261) treść intencjonalna reprezentacji pojęciowych jest determinowana „receptorowo", czyli na podstawie współzmienności reprezentacji z występowaniem egzemplarzy danej kategorii. W swojej pracy odrzucam tę konkretną tezę.

aktywacją tychże systemów w taki sposób, że pełnią one nowe, wyższe (pojęciowe) funkcje poznawcze.

Barsalou (1999) formułuje TSP, wprowadzając pojęcia „symulatora" oraz „symulacji". Zacznijmy od omówienia symulatorów. Według Barsalou każdej interakcji z obiektem należącym do określonej kategorii towarzyszy pewien wzór aktywności neuronalnej w systemach percepcyjnych i systemie motorycznym. Zgodnie z TSP ten wzór jest rejestrowany i zapamiętywany w obszarach asocjacyjnych mózgu – tak zwanych strefach konwergencji (*convergence zones*); por.: Damasio 1989; Meyer, Damasio 2009. Co więcej, wzory aktywności percepcyjno-motorycznej są tam integrowane ze sobą, tworząc multimodalny i motoryczny „profil" danego obiektu, obejmujący własności związane z modalnością wzrokową, dotykową, słuchową, introcepcją, jak również sekwencje motoryczne towarzyszące interakcjom z tym obiektem. Ważnym elementem TSP pozostaje twierdzenie, że interakcjom z obiektami o tej samej przynależności kategorialnej towarzyszą podobne czy zbliżone wzory neuronalnej aktywności percepcyjnej i motorycznej. Ze względu na zachodzące między nimi podobieństwo – wzory te są rejestrowane w zbliżonych rejonach obszarów asocjacyjnych. Właśnie w taki sposób powstaje po pewnym czasie multimodalny i motoryczny ślad pamięciowy *kategorii* – symulator (Barsalou 1999). Każdy symulator „składuje" neuronalną pamięć percepcyjnych i motorycznych interakcji z obiektami należącymi do danej kategorii. Na przykład symulator kategorii ROWER to przechowywany w pamięci długotrwałej ślad pamięciowy percepcyjnej i motorycznej aktywności neuronalnej towarzyszącej interakcjom z obiektami należącymi do tej kategorii. Symulator ten składuje pamięć (między innymi) o tym, jak rowery wyglądają, jakie wydają dźwięki, jak są praktycznie używane i jakie doznania interoceptywne towarzyszą ich wykorzystywaniu.

Zgodnie z TSP symulatory mogą „odgórnie" reaktywować obszary percepcyjne i motoryczne, przy czym wzorce tak wywołanej aktywności są podobne do wzorców aktywności, która mogłaby towarzyszyć rzeczywistej interakcji z egzemplarzem danej kategorii (na przykład konkretnym rowerem). Właśnie taką opartą na symulatorze reaktywację percepcyjno-motoryczną Barsalou nazywa „sy-

mulacją" (1999). Symulacje są przeprowadzane w pamięci roboczej i nigdy nie odzwierciedlają całej zawartości odpowiadającego im symulatora. Są one zawsze selektywne i fragmentaryczne. Co więcej – mogą różnić się w różnych kontekstach. Na przykład symulacja interakcji z rowerem wykorzystywana przy udzielaniu odpowiedzi na pytanie o to, czy rowery mają koła, może różnić się do symulacji używanej w celu udzielenia odpowiedzi na pytanie o to, czy rowery są szybsze od samochodów. Warto też zaznaczyć, że w perspektywie TSP symulacje nie są holistyczne, lecz złożone. Podczas percepcyjno-motorycznej symulacji interakcji z rowerem mechanizmy uwagowe mogą wyizolować pewne fragmenty i utworzyć na tej podstawie odrębny symulator kategorii KOŁO ZĘBATE. Analogicznie, jest możliwe integrowanie symulatorów i tworzenie na tej podstawie nowych, złożonych symulacji, które nie muszą odpowiadać żadnej wcześniejszej percepcji.

Według Barsalou (1999) symulatory i oparte na nich symulacje mogą realizować wszystkie funkcje poznawcze, które tradycyjnie przypisuje się reprezentacjom pojęciowym[50]. Autor ten argumentuje, że na gruncie TSP można wyjaśnić takie zdolności, jak predykacja, przeprowadzanie rozumowań czy kategoryzacja percepcyjna. Pokazuje on także, że TSP nie stoi w sprzeczności z ideą, iż myślenie pojęciowe jest kompozycyjne i produktywne. Wreszcie inni autorzy starają się dodatkowo pokazać, że TSP może być zaaplikowana nie tylko do pojęć obiektów fizycznych, ale także do pojęć stanów

[50] Weiskopf (2007) zauważa, że TSP – nazywana przez niego „empiryzmem pojęciowym" – może być rozumiana na kilka sposobów. Autor ten odróżnia: (1) mocny globalny empiryzm pojęciowy („wszystkie myśli są w całości złożone z symulacji percepcyjno-motorycznych"); (2) słaby globalny empiryzm pojęciowy („wszystkie myśli są częściowo złożone z symulacji percepcyjno-motorycznych"); (3) mocny lokalny empiryzm pojęciowy („niektóre myśli są w całości złożone z symulacji percepcyjno-motorycznych") oraz (4) słaby lokalny empiryzm pojęciowy („niektóre myśli są w części złożone z symulacji percepcyjno-motorycznych"). Choć prowadzona tu dyskusja pozostaje neutralna ze względu na to, która z wymienionych wersji TSP jest prawdziwa, warto mieć na uwadze, że TSP nie musi być wcale rozumiana jako koncepcja radykalna, zgodnie z którą cała aktywność pojęciowa sprowadza się do symulacji percepcyjno-motorycznych.

mentalnych (por.: Niedenthal, Winkielman, Mondillon, Vermeulen 2009; Gładziejewski 2013b) czy pojęć abstrakcyjnych i esencjalistycznych (por. Prinz 2002: 263–282).

Co bardzo istotne, TSP dysponuje znaczącym wsparciem empirycznym[51]. Choć pełen przegląd wyników eksperymentalnych nie jest tu możliwy, warto zaznaczyć, że na korzyść TSP wydają się świadczyć zarówno badania behawioralne, jak i badania wykorzystujące neuroobrazowanie. Jeśli chodzi o badania o charakterze czysto behawioralnym, okazuje się, iż (1) osoby proszone o generowanie list własności charakteryzujących egzemplarze danej kategorii wymieniają częściej i szybciej te własności, które byłyby najłatwiej dostępne w ramach (postulowanej) symulacji percepcyjno-motorycznej (na przykład szybciej i częściej wymieniają „posiadanie pestek" dla kategorii „połówka arbuza" niż dla kategorii „arbuz"; por.: Barsalou, Solomon, Wu 1999; Wu, Barsalou 2009); (2) osoby realizujące zadanie polegające na weryfikowaniu, czy egzemplarze danej kategorii mają określoną własność, wykazują – w zgodzie z predykcjami TSP – dłuższy czas reakcji, weryfikując własności dostępne za pomocą różnych modalności zmysłowych (na przykład weryfikując parę „pies–szczeka" bezpośrednio po uprzedniej weryfikacji pary „pies–purpurowy"), niż przy weryfikowaniu własności dostępnych za pomocą jednej modalności (na przykład weryfikując parę „pies–wyje" po weryfikacji pary „pies–szczeka"; por. Pecher, Zeelenberg, Barsalou 2003). Jeśli zaś chodzi o badania z wykorzystaniem neuroobrazowania, to pokazują one, iż realizowaniu zadań eksperymentalnych wymagających wykorzystania reprezentacji pojęciowych towarzyszy aktywność zmysłowo specyficznych, percepcyjnych obszarów mózgu oraz obszarów motorycznych (por.: Martin 2007; Simmons, Ramjee, Beauchamp et al. 2007).

Sądzę, że opisany przez Barsalou mechanizm złożony z symulatorów i symulacji stanowi pewną formę MKM. Koncepcja symboli percepcyjnych może być zinterpretowana jako postulująca, że ope-

[51] Warto mieć jednak na uwadze, że krytycy TSP starają się pokazać, iż często nie sposób nadać całkiem jednoznaczną interpretację danym przytaczanym na rzecz tej teorii. Na fakt ten zwraca uwagę przede wszystkim Machery (2007; por. jednak odpowiedź na jego zarzuty w: Prinz 2010).

rowanie pojęciowymi reprezentacjami kategorii opiera się na MKM. Warto jednak *explicite* zaznaczyć, jak dokładnie rozumiem to twierdzenie, tym bardziej że sama TSP jest podatna na filozoficznie naiwne interpretacje.

Gdybyśmy trzymali się sformułowań często używanych przez samego Barsalou (1999, 2009), należałoby stwierdzić, że procesowi symulacji postulowanemu przez TSP podlegają same kategorie albo możliwe egzemplarze kategorii. Sądzę jednak, że takie postawienie sprawy jest błędem. Proponuję uznać, że percepcyjno-motoryczna symulacja pozwalająca dla przykładu na wyciągnięcie wniosku, iż koty miauczą, nie jest wcale symulacją kota czy kategorii KOT. W świetle (poprawnie zinterpretowanej) TSP mózgi nie symulują kotów ani żadnych innych *obiektów w świecie*. Co zatem podlega symulacji? Jak się wydaje, postulowane przez Barsalou symulacje polegają na modelowaniu *wzorców aktywności neuronalnej kontrolującej/regulującej interakcje* z różnymi obiektami w świecie (na przykład kotami). W świetle takiej interpretacji TSP głosi, że mózg, realizując funkcje pojęciowe, wykorzystuje symulacje procesów neuronalnych czy neuroobliczeniowych kontrolujących bezpośrednie interakcje z obiektami. Ośrodkowy układ nerwowy nie zajmuje się symulacją samych egzemplarzy/kategorii; symulacje polegają na *reaktywowaniu aktywności neuronalnej* (neuroobliczeniowej). Symulowane są nie kategorie czy egzemplarze, lecz przebieg interakcji z egzemplarzami (przebieg neuronalnych procesów kontrolujących te interakcje). Z kolei konsumentami takich symulacji są komponenty systemu poznawczego wyspecjalizowane w pełnieniu typowo konceptualnych funkcji poznawczych. To właśnie one funkcjonalnie wykorzystują podobieństwo zachodzące między wewnętrznymi symulacjami interakcji a samymi interakcjami. Na przykład kiedy pytamy kogoś o to, czy koty miauczą, to (w olbrzymim uproszczeniu): (1) w mózgu tej osoby symulator kategorii KOT generuje symulację interakcji z egzemplarzem kota, obejmującą również słuchowe aspekty takiej interakcji (obszar neuronalny dokonujący symulacji jest tu nośnikiem reprezentacji); (2) komponent (układ komponentów) zaangażowany w funkcje związane z użyciem języka naturalnego (także w kontrolowanie komunikacji werbalnej) na podsta-

wie przebiegu symulacji generuje werbalną odpowiedź twierdzącą (komponent ten jest konsumentem). Konsument generuje tu poprawną odpowiedź dzięki temu, że aktywność nośnika wiernie odzwierciedla pewien aspekt interakcji z kotami.

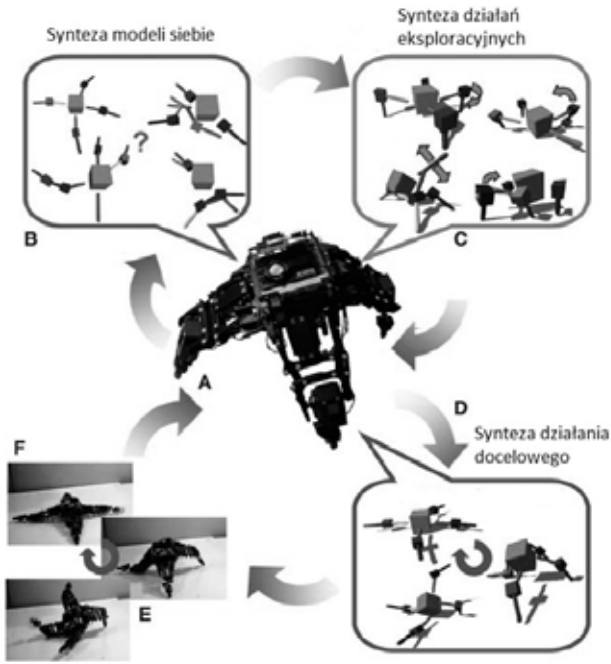
Na jakiej dokładnie zasadzie symulowanie (modelowanie) swojej własnej aktywności pozwala systemowi poznawczemu na realizowanie funkcji pojęciowych? Pewną drogę w kierunku rozwiązania tego problemu wskazuje sugestia, że da się połączyć koncepcję symboli percepcyjnych z omówioną wyżej teorią emulacji (Pezzulo 2011). Zgodnie z tą propozycją postulowane przez Barsalou symulacje są w istocie wynikiem aktywności emulatorów. Posługiwanie się pojęciami opiera się na wykorzystaniu emulatorów całkowicie *off-line*, w oderwaniu od kontekstu bezpośredniej interakcji z obiektami w świecie. Inaczej mówiąc, korzystanie z pojęć (reprezentacji pojęciowych) opiera się na aktywności *mocno oderwanych emulatorów*. Aktywność ta byłaby wykorzystywana („konsumowana") przez komponenty mózgu odpowiedzialne funkcjonalnie za wyższe funkcje poznawcze[52]. Pozostawiam tę propozycję jako otwartą, choć obecnie jedynie czysto spekulatywną możliwość.

d) MKM w robotyce poznawczej: samomodelujący się robot Bongarda, Zykova i Lipsona

Robotyka poznawcza jest często uznawana – między innymi za sprawą wpływowych prac Rodneya Brooksa (1991) – za „bastion" antyreprezentacjonizmu we współczesnej kognitywistyce. Jednak również ten obszar nauk kognitywnych nie jest pozbawiony odwołań do MKM. Dobrym przykładem obecności modeli wewnętrznych w robotyce poznawczej jest samomodelujący się robot zaprojektowany

[52] Por. idące w podobnym kierunku, choć bardzo ostrożne stwierdzenie samego Grusha: „Niech mi będzie wolno tylko powiedzieć, że uważam za użyteczne uznanie artykułowanych emulatorów za pojęciowe, w takim sensie, że artykulanty [komponenty emulatorów – P. G.] mają wiele cech pojęć. Nie będę jednak kontynuował tych rozważań. Problem tego, czym są pojęcia, jest skomplikowany i jeżeli mam być szczery – nie leży w centrum moich zainteresowań" (2010: 200). Co warte wspomnienia, w przytaczanym wywiadzie Grush wyraża też przekonanie, że emulatory odgrywają rolę we wnioskowaniach.

przez Josha Bongarda, Victora Zykova i Hoda Lipsona (2006; por. Adami 2006).



Rysunek 8. Samomodelujący się robot skonstruowany przez Josha Bongarda, Victora Zykova i Hoda Lipsona. Źródło: Bongard, Zykov, Lipson 2006: 1119

Robot skonstruowany przez Bongarda i współpracowników ma kształt rozgwiazdy. Wyposażono go w dwa czujniki nachylenia ciała (*tilt sensors*) oraz cztery kończyny połączone za pomocą ośmiu stawów wyposażonych w czujniki kątów zgięcia (por. rysunek 8). Co istotne dla naszych celów, robot ten porusza się w świecie dzięki procesowi ciągłego, autonomicznego modelowania samego siebie. Jest on zaprojektowany tak, by tworzyć obliczeniowe modele morfologii swojego własnego ciała, a następnie wykorzystywać te mo-

dele w procesach związanych z kontrolą motoryczną i planowaniem działań. Tworzenie modelu siebie jest inicjowane wykonaniem szeregu losowych ruchów, w trakcie których rejestruje się aktywność wszystkich czujników (por. rysunek 8-A). Następnie dochodzi do procesu generowania i testowania modeli ciała własnego. Najpierw odpowiedni komponent robota generuje 15 potencjalnych modeli, z których każdy jest spójny z zarejestrowanymi wcześniej wzorami aktywności sensorycznej (por. rysunek 8-B). Bongard i współpracownicy (2006) nazywają ten etap „syntezą modeli siebie". Następnie – w procesie nazywanym „syntezą działań eksploracyjnych" (por. rysunek 8-C) – zostaje wybrane potencjalne działanie, którego wykonanie mogłoby pozwolić na selektywną weryfikację wygenerowanych modeli. Zostaje wyselekcjonowane działanie, które w świetle poszczególnych wygenerowanych uprzednio modeli powinno wywołać różne wzorce aktywności sensorycznej, pozwalając tym samym na porównawczą weryfikację mocy predykcyjnej tych modeli. Następnie robot wykonuje to działanie. Teraz cały cykl jest powtarzany – tym razem jednak nie całkowicie „od zera", lecz biorąc pod uwagę wyniki wcześniejszego testu. Dochodzi zatem do syntezy kolejnych modeli, a następnie do syntezy działań eksploracyjnych oraz wykonania kolejnego działania testującego dostępne modele. Cykl ten powtarza się 16 razy. W wyniku całego procesu jest selekcjonowany jeden, najlepszy model, który zostaje wykorzystany przez komponent dokonujący tak zwanej syntezy działania docelowego (por. rysunek 8-D), które to (docelowe) działanie zostaje następnie wykonane (por. rysunek 8-E). Cały opisany proces może zostać powtórzony (por. rysunek 8-F), w zależności od potrzeb, na jednym z dwóch etapów: etapie syntezy modeli (aby dodatkowo udoskonalić stworzony wcześniej model) albo na etapie syntezy działania docelowego (aby stworzyć nowe schematy działania).

Ważną cechą robota zaprojektowanego przez Bongarda i współpracowników jest jego zdolność do behawioralnego adaptowania się do zmian w morfologii ciała własnego. Po jego uszkodzeniu (na przykład po utracie fragmentu kończyny) robot będzie w stanie wykryć predykcyjną nieadekwatność posiadanego modelu (który został wytworzony przed uszkodzeniem) i ponownie zainicjować na

tej podstawie proces syntezy modeli. Wykorzystując posiadany dotychczas model jako punkt wyjścia, proces generowania i testowania doprowadza do powstania nowego, dopasowanego do aktualnej morfologii robota modelu. W ten właśnie sposób proces samomodelowania pozwala na plastyczne adaptowanie się robota do zmian zachodzących w jego sztucznym ciele.

Zanim przejdziemy dalej, należy poruszyć dwie kwestie. Pierwsza dotyczy relacji między koncepcją MKM a opisanym tu robotem. W jakim sensie mechanizm działania robota skonstruowanego przez Bongarda i współpracowników stanowi formę MKM? Zwróćmy uwagę, że model ciała generowany w opisanym wyżej procesie przypomina to, co Grush nazwał emulatorem artykułowanym. Jest on S-reprezentacją ciała własnego robota, która ma docelowo odzwierciedlać kształt tego ciała oraz możliwe interakcje między kończynami. Model okazuje się tym bardziej funkcjonalny dla robota, im lepiej odzwierciedla on morfologię jego ciała (jest strukturalnie podobny do tego ciała). Konsumenta modelu stanowi zaś komponent odpowiedzialny za syntezę działania docelowego. Wykorzystuje on surogatywnie wewnętrzny model zamiast samego ciała, aby zaplanować pewne działanie docelowe. Innymi słowy, komponent ten wykorzystuje wierność modelu – jest jej funkcjonalnym „beneficjentem" – względem rzeczywistej morfologii ciała, aby wygenerować odpowiedni schemat działania.

Druga uwaga, która musi zostać tu poczyniona, wiąże się z faktem, że robot Bongarda i współpracowników niezbyt dobrze wpisuje się w Ramseya odróżnienie na klasyczną i nieklasyczną kognitywistykę. Z jednej strony korzysta on z klasycznej, symbolicznej architektury obliczeniowej, a nie na przykład z sieci koneksyjnej. W ten właśnie sposób wpisuje się on w GOFAI. Z drugiej jednak strony omawiana praca nie do końca może być sklasyfikowana jako należąca do klasycznej kognitywistyki, ponieważ (1) kładąc nacisk na znaczenie ciała własnego w poznaniu (zwróćmy uwagę na rolę działań eksploracyjnych przy tworzeniu modeli), a także na rolę reprezentacji (modeli) ciała i wykorzystanie takich reprezentacji w dokonywanych *on-line* interakcjach ze światem, praca Bongarda i współpracowników wydaje się zgodna z duchem „nowej", ucieleśnionej

kognitywistyki; (2) Bongard i współpracownicy (2006) sami traktują użycie klasycznej, symbolicznej architektury obliczeniowej jako rozwiązanie prowizoryczne, które docelowo powinno zostać porzucone i zastąpione bardziej realistyczną biologicznie wizją obliczeniowej architektury poznania[53]. Z tych dwóch powodów bardzo trudno przyjąć, że opisywany tu robot stanowi zaledwie intelektualny „relikt" kognitywistyki. Dlatego też uznaję za zasadne wykorzystanie pracy Bongarda i współpracowników w roli kontrprzykładu dla twierdzenia Ramseya o dezaktualizacji reprezentacji jako modeli mentalnych w nieklasycznej kognitywistyce.

e) MKM w neuronauce obliczeniowej: system okulomotoryczny

Neuronauka obliczeniowa wspiera się na założeniu, że mózg to nie tylko system, który można w użyteczny sposób obliczeniowo opisywać czy przewidywać – pod tym względem nie różniłby się on wszakże chociażby od procesów pogodowych – ale także system, którego działanie w jakimś nietrywialnym sensie opiera się na wykonywaniu obliczeń (por.: Piccinini 2007a, 2007b; Shagrir 2010; Miłkowski 2013). Neuronaukowcy obliczeniowi zajmują się stworzeniem biologicznie realistycznej koncepcji tego, w jaki dokładnie sposób ośrodkowy układ nerwowy wykonuje obliczenia. Angażują się oni również w formułowanie inspirowanych działaniem mózgu obliczeniowych wyjaśnień poszczególnych funkcji poznawczych. Jak się okazuje, także w tym obszarze kognitywistyki jest obecne pojęcie reprezentacji jako modeli mentalnych (por. Shagrir 2010; 2012). Ilustrację takiego stanu rzeczy przedstawia Oron Shagrir (2012), powołując się na badania nad neuroobliczeniowymi podstawami kontroli okulomotorycznej (kontroli ruchów gałek ocznych przez mózg).

[53] Jak to ujmują ci autorzy: „Choć jest mało prawdopodobne, że organizmy dysponują jawnymi [*explicit*] modelami w rodzaju tych opisanych tutaj, to proponowana przez nas metoda może naświetlić naturę nieznanych procesów, za pomocą których organizmy tworzą i uaktualniają modele samych siebie w mózgu, jak również naświetlić to, jak organizmy używają w tym celu sygnałów sensomotorycznych, jaką formę przyjmują takie modele oraz na czym polega przydatność konkurencji między wieloma alternatywnymi modelami" (Bongard, Zykov, Lipson 2006: 1121).

Prace, na które powołuje się Shagrir (2012), dotyczą zdolności systemu okulomotorycznego do sakadowych i międzysakadowych horyzontalnych fiksacji gałek ocznych (utrzymywania gałek w horyzontalnie nieruchomej pozycji). Kontrolując ruchy oczu, mózg musi ustawicznie śledzić aktualne położenie gałek, utrzymywać to położenie, a także zmieniać je, kiedy zachodzi taka potrzeba. Wyniki badań każą sądzić, że ośrodkowy układ nerwowy realizuje te funkcje dzięki korzystaniu z wyspecjalizowanej w tym celu pamięci krótkotrwałej. Pamięć ta ma być implementowana przez sieć neuronową stanowiącą multistabilny system dynamiczny. Obliczeniowy model tej sieci nie ma warstwy wejściowej i wyjściowej, a wszystkie neurony wchodzące w jego skład są wzajemnie połączone. W każdym momencie sieć zajmuje pewne miejsce w przestrzeni stanów, której poszczególne punkty odpowiadają możliwym globalnym stanom sieci. Ewolucja sieci w czasie stanowi trajektorię w tej przestrzeni. Pojawienie się nowego bodźca w polu wzrokowym sprawia, że sieć jest wytrącana z jednego stabilnego stanu (stałego wzorca aktywności neuronalnej) i zmierza ku innemu stabilnemu stanowi. Aktywność sieci ma realizować (obliczać) funkcję matematyczną przetwarzającą sygnały kodujące prędkość poruszania się gałek ocznych (*eye-velocity*) na sygnały motoryczne zmieniające położenie gałek z jednego punktu fiksacji w inny. Znaczenie omawianej sieci jako pamięci polega na tym, że pozwala ona na utrzymanie gałek ocznych w jednym położeniu nawet wtedy, gdy sygnał wejściowy już zniknął. W realnych mózgach uszkodzenie obszarów implementujących tę pamięć skutkuje tym, że oczy zaczynają dryfować w niekontrolowany sposób.

Istotne dla tego wywodu jest to, jak opisanej sieci udaje się pełnić funkcję krótkotrwałej pamięci położenia gałek ocznych. Otóż Shagrir (2012) zwraca uwagę na rolę, jaką odgrywa podobieństwo zachodzące między strukturą opisującej tę sieć (jej ewolucję w czasie) przestrzeni stanów a strukturą możliwych pozycji gałek ocznych. Okazuje się bowiem, że metryczna struktura owej przestrzeni stanów odzwierciedla strukturę przestrzenną możliwych położeń oczu. Każdy możliwy stabilny stan *Si* sieci odpowiada jednemu możliwemu położeniu gałek ocznych – *Ei*. Im bliżej siebie znajdują się dwa

możliwe stany sieci (rozumiane jako punkty w przestrzeni stanów), tym bliżej znajdą się odpowiadające im położenia gałek ocznych. Kiedy sieć przechodzi ze stabilnego stanu *Si* w stabilny stan *Sj*, oczy przechodzą z odpowiadającego *Si* stabilnego horyzontalnego położenia (fiksacji) *Ei* w odpowiadające *Sj* stabilne horyzontalne położenie *Ej*. Trajektoria ewolucji sieci w przestrzeni stanów dynamicznie odzwierciedla horyzontalne ruchy gałek ocznych. To podobieństwo strukturalne nie jest jedynie epifenomenem. Pełni ono zasadniczą rolę w wyjaśnianiu tego, jak system okulomotoryczny steruje horyzontalnymi ruchami gałek ocznych. System okulomotoryczny *wykorzystuje* bowiem to podobieństwo. Jak to ujmuje Shagrir:

> Przestrzeń stanów sieci odzwierciedla przestrzeń pozycji oczu. [...] Przestrzeń stanów sieci może być postrzegana jako *mapa*, w której linia atraktora odpowiada przestrzeni pozycji oczu. [...] W rzeczy samej system okulomotoryczny wykorzystuje tę mapę przy wykonywaniu różnych zadań neuropoznawczych. W omawianym tu kontekście neurony motoryczne „odczytują" obecny stan sieci, aby skierować oczy w kierunku nowego położenia i utrzymać je w tym położeniu. [...] [Sieć] składa się z konkretnych stanów wewnętrznych, z którymi system okulomotoryczny może się „konsultować" podczas rozwiązywania problemów (Shagrir 2012: 533).

Mamy tu zatem do czynienia z dość klarownym przykładem wyjaśnienia pewnego fenomenu poznawczego za pomocą mechanizmu korzystającego z wewnętrznego modelu. Sieć implementująca krótkotrwałą pamięć horyzontalnych położeń gałek ocznych sprawuje funkcję *modelu* (jest nośnikiem modelu). Dzieje się tak z dwóch powodów. Po pierwsze, przestrzeń stanów sieci wykazuje strukturalne podobieństwo do układu czy struktury pozycji gałek ocznych (stanowiących przedmiot reprezentacji). Po drugie, wymienione podobieństwo jest konsumowane przez system okulomotoryczny w celu podejmowania „decyzji" motorycznych, to znaczy do odpowiedniego sterowania ruchami gałek ocznych. Mózg kontroluje zatem horyzontalny ruch gałek ocznych, wykorzystując w tym celu konsumowany model. Co bardzo istotne, według Shagrira (2012) takie wykorzystanie S-reprezentacji jako narzędzia eksplanacyjnego nie

jest jedynie odosobnionym przypadkiem, lecz praktyką powszechnie spotykaną w neuronauce obliczeniowej.

Przedstawiony przegląd nie jest z pewnością wyczerpujący. Zastosowanie pojęcia reprezentacji jako modeli mentalnych w nieklasycznej kognitywistyce zdecydowanie wykracza poza wymienione przypadki. Jako kolejne (potencjalne) ilustracje, można by wymienić chociażby: (1) koncepcję kory mózgowej SINBaD Dana Rydera (2004); (2) opartą na samomodelowaniu koncepcję samoświadomości Thomasa Metzingera (2003); (3) symulacyjną teorię czytania umysłów w odmianie rozwijanej przez Alvina Goldmana (2006; por. krytyczne omówienie roli pojęcia modelowania i symulacji w tej koncepcji, zawarte w: Herschbach 2012); (4) teorię przestrzeni pojęciowych Petera Gardenförsa (2000); (5) koncepcje pamięci epizodycznej i prospekcji („mentalnej podróży w czasie") oparte na idei, że zdolności te wykorzystują mechanizm oparty na symulacji percepcyjno-motorycznej podobnej do tej, którą postuluje Barsalou (por.: Kent, Lamberts 2008; Danker, Anderson 2010; Schacter, Addis, Hassabis et al. 2012)[54]. Na osobną wzmiankę zasługuje także szybko zyskująca obecnie na znaczeniu teoria, zgodnie z którą działanie mózgu jest oparte na kodowaniu predykcyjnym (Friston, Stephan 2007; Clark 2013; Hohwy 2013). Według tej koncepcji aktywność mózgu polega na wykorzystywaniu *modelu generatywnego* środowiska w celu zredukowania błędu predykcyjnego, stanowiącego różnicę między rzeczywistymi a przewidywanymi wzorcami pobudzenia sensorycznego[55]. Podejrzewam, że jest dopuszczalna interpretacja tej teorii, zgodnie

[54] Zakładam tu rzecz jasna, że każda z wymienionych teorii może zostać pojęciowo uzgodniona z MKM (można w niej wyróżnić nośnik modelu, jego konsumenta/konsumentów oraz przedmiot reprezentacji). Sądzę, że w każdym przypadku takie uzgodnienie jest wykonalne, choć nie będę się tu podejmować tego zadania.

[55] Teoria odwołująca się do kodowania predykcyjnego wykazuje uderzające podobieństwo do teorii emulacji: obie koncepcje kładą nacisk na rolę predykcji w działaniu mózgu oraz podkreślają fundamentalne znaczenie poznawcze, jakie ma dla mózgu określenie różnicy między wynikami wewnętrznej predykcji a (częściowo) „niespodziewanymi" sygnałami napływającymi ze środowiska lub ciała (Clark 2013).

z którą model generatywny stanowi pełnoprawny (konsumowany) „model" w sensie przyjmowanym w koncepcji MKM (por. Gładziejewski, w druku). Model generatywny miałby w takim ujęciu odzwierciedlać strukturę statystyczną/przyczynową świata, generując dzięki temu predykcje sensoryczne, które byłyby konsumowane przez komponenty mózgu odpowiedzialne za percepcję oraz działanie.

Istnienie oraz znacząca rola wszystkich wymienionych w tym podrozdziale koncepcji – nie mówimy wszakże o propozycjach stanowiących jakiś „egzotyczny" margines nauk kognitywnych – każe wyciągnąć wniosek, że *współczesna (nieklasyczna) kognitywistyka jest w znaczącym stopniu reprezentacjonistyczna*. Znacząca część jak najbardziej aktualnych, posiadających empiryczne wsparcie wyjaśnień czy teorii z zakresu kognitywistyki odwołuje się bowiem do struktur, które spełniają przedstawioną w sekcji 4.2.1 charakterystykę konsumowanych modeli mentalnych. Inaczej mówiąc, strategia eksplanacyjna polegająca na odwołaniu się do MKM jako eksplanansu danego zjawiska jest, a nie jedynie była, stosunkowo powszechna w naukach kognitywnych. Co więcej, zastosowanie tej strategii nie ogranicza się do jakiegoś pojedynczego nieklasycznego podejścia, ale wydaje się przenikać wszystkie lub niemal wszystkie rejony współczesnej kognitywistyki. Na ile ta ostatnia daje nam rzeczywisty wgląd w działanie systemów poznawczych, na tyle uzasadnione będzie twierdzenie, iż systemy te są w jakimś zakresie reprezentacyjne.

Powyższa konstatacja wymaga pewnego doprecyzowania. W rozdziale 1 (sekcja 1.1.3) została odróżniona globalna wersja reprezentacjonizmu na poziomie przedmiotowym oraz lokalna wersja tego stanowiska. O ile to pierwsze wymienione stanowisko opiera się na twierdzeniu, że system poznawczy jako taki w całości jest systemem reprezentacyjnym, o tyle to drugie zostaje zrelatywizowane do jakiegoś pojedynczego, wyjaśnianego fenomenu kognitywnego. Otóż diagnozując współczesną kognitywistykę jako „w znaczącym stopniu" reprezentacjonistyczną, moją intencją nie jest postawienie tezy o triumfie globalnej wersji reprezentacjonizmu przedmiotowego. Twierdzę raczej, że są podstawy, by sądzić, iż reprezentacjonizm triumfuje obecnie – lub przynajmniej stanowi liczącą się, relewant-

ną opcję teoretyczną – na wielu lokalnych frontach[56]. Dopuszczam istnienie zjawisk poznawczych, które będą z powodzeniem wyjaśniane za pomocą mechanizmów niewykorzystujących reprezentacji. Nie ulega wątpliwości, że współczesna kognitywistyka pozostaje do pewnego stopnia antyreprezentacjonistyczna. Twierdzę jednak zarazem, iż pojęcie wewnętrznych, konsumowanych modeli wykorzystuje się współcześnie na tyle często w znaczących, dobrze ugruntowanych wyjaśnieniach poszczególnych zjawisk poznawczych, że Ramseyowskie twierdzenie o upadku reprezentacjonizmu w nieklasycznej kognitywistyce powinno zostać uznane za zupełnie nieuzasadnione. Reprezentacjonizm nadal żyje w naukach kognitywnych i ma się całkiem dobrze – nawet jeśli nie stanowi już niepodważalnego dogmatu.

[56] Dopuszczam co prawda możliwość, iż rozwój nauk kognitywnych ostatecznie doprowadzi do unifikacji teoretycznej kognitywistyki „pod banderą" reprezentacjonizmu. Clark (2013) perspektywę unifikacji upatruje chociażby w teorii mózgu jako systemu korzystającego z modeli generatywnych środowiska. Gdyby tego rodzaju scenariusz się ziścił, mogłoby to oznaczać triumf globalnego reprezentacjonizmu (zakładając, że modele generatywne rzeczywiście stanowią formę modeli mentalnych w przyjmowanym tu rozumieniu). Nie chcę tu jednak wdawać się w spekulacje dotyczące przyszłości kognitywistyki, skupiając się raczej na jej faktycznym stanie obecnym.

ROZDZIAŁ 5

Reprezentacjonizm w kognitywistyce a problem naturalizacji intencjonalności

5.1. Naturalizowanie intencjonalności i dystynkcja osobowe–subosobowe

5.1.1. Dwa obrazy Sellarsa a osobowy i subosobowy poziom wyjaśniania

Przypomnijmy sobie teraz omówione we wstępie tej książki Sellarsowskie rozróżnienie na naukowy i manifestujący się obraz świata. Pierwszy, naukowy obraz zawiera się w najlepiej uzasadnionych teoriach, modelach i wyjaśnieniach formułowanych przez przedstawicieli nauk szczegółowych. Drugi, manifestujący się obraz to ten, który jest obecny w codziennych interakcjach ludzi z zamieszkiwanym przez nich światem. Pierwszy obraz przedstawia świat jako „atomy w próżni", drugi zaś przedstawia uniwersum posiadające „ludzkie" oblicze: świat wartości, instytucji społecznych, historii, ale także świat wypełniony podmiotami stanów mentalnych. Przypomnijmy sobie też Sellarsowski projekt filozofii jako dyscypliny, której cel stanowi uzgodnienie obu wymienionych obrazów, pokazanie, że jest możliwe połączenie ich w ramach jednolitej, niesprzecznej, synoptycznej wizji Wszechświata.

Zaproponowana tu teoria mechanizmów reprezentacyjnych miała stanowić odpowiedź na pytanie o naturę wyjaśniania reprezentacyjnego w naukach kognitywnych. Kognitywistyka stanowi jednak element naukowego obrazu świata. Do tej pory poruszałem się w tej pracy jedynie w obrębie tego obrazu. Nie podjąłem się jak dotąd realizacji jednego z celów tej książki, który został zaznaczony we wstępie oraz rozdziale 1. Nie spróbowałem przerzucić mostu między naukowym obrazem umysłu a obrazem manifestującym

się. Zadaniem tego ostatniego rozdziału jest właśnie podjęcie – wykorzystując w tym celu mechanistyczny model wyjaśniania – próby zbudowania takiego mostu.

Jak zostało wspomniane we wstępie, obraz manifestujący się przedstawia świat jako wypełniony istotami działającymi na podstawie posiadanych stanów mentalnych. Przynajmniej niektóre z tych stanów – postawy propozycjonalne – cechują się intencjonalnością. Postawami propozycjonalnymi są między innymi pragnienia, nadzieje, intencje, przekonania czy wątpliwości. Wszystkie one mają treść intencjonalną, są o czymś czy też mają jakieś warunki prawdziwości lub poprawności (spełnienia). Codzienne ludzkie interakcje opierają się w większym lub mniejszym stopniu[1] na atrybuowaniu sobie nawzajem postaw propozycjonalnych oraz wyjaśnianiu i przewidywaniu na tej podstawie działań.

Jak zostało już w opisane rozdziale 1 (podrozdział 1.2), niektórzy filozofowie chcą uzgodnić dwa obrazy umysłu czy poznania – ten zawarty w psychologii potocznej oraz ten dostarczany przez kognitywistykę – angażując się w projekt naturalizacji intencjonalności. Mówiąc w zarysie, są oni zainteresowani udzieleniem odpowiedzi na pytanie o to, jak naturalistycznie pojmowane systemy poznawcze mogą być podmiotami postaw propozycjonalnych. Chcą oni zaproponować filozoficznie satysfakcjonującą koncepcję tego, jak system fizyczny opisywany i wyjaśniany w sposób odwołujący się jedynie do *stricte* naturalnych własności może być systemem działającym na podstawie pragnień, przekonań i innych postaw propozycjonalnych. Trzeba wyraźnie zaznaczyć, że zasadniczym celem tego rozdziału nie jest rozwikłanie problemu naturalizacji intencjonalności (choć pewne sugestie w tej sprawie zostaną wyrażone w podrozdziale 5.3). Chcę raczej krytycznie przemyśleć filozoficzne podstawy tego projektu i niektóre presupozycje, na których tle jest on rozwijany. Zamiast naturalizować intencjonalność, chcę zająć się następującymi pytaniami: jaka jest relacja między reprezentacyjnymi wyjaśnienia-

[1] To, jak często ludzie rzeczywiście posługują się tego rodzaju kategoriami w toku codziennych interakcji społecznych, stanowi w ostatnich latach przedmiot dyskusji (por. m.in.: Zahavi 2006; Ratcliffe 2007; Herschbach 2008; Spaulding 2010), w której nie chcę tu jednak zajmować określonej pozycji.

mi formułowanymi w ramach kognitywistyki a wyjaśnieniami odwołującymi się do postaw propozycjonalnych? Czy rozstrzygnięcia dotyczące statusu eksplanacyjnego reprezentacji w kognitywistyce niosą ze sobą jakieś konsekwencje dla statusu wyjaśnień w kategoriach postaw propozycjonalnych? Czy aby projekt naturalizacji intencjonalności się powiódł, musimy pokazać, że wyjaśnienia z zakresu kognitywistyki w jakimś sensie „rehabilitują" albo „naturalizują" wyjaśnienia formułowane w ramach obrazu manifestującego się? Czy kiedy wyjaśniamy czyjeś działania za pomocą przekonań i pragnień, to poprawność naszej praktyki eksplanacyjnej zależy w jakiś sposób od faktów dotyczących strukturalno-funkcjonalnej organizacji systemu poznawczego? Mówiąc zatem ogólnie, zamiast angażować się po prostu w projekt naturalizacji – pragnę tu przede wszystkim przemyśleć relację między tym właśnie projektem a podejmowaną tu przeze mnie do tej pory kwestią eksplanacyjnej użyteczności reprezentacji w kognitywistyce. Chcę pokazać, w jaki sposób ta relacja jest na ogół rozumiana w analitycznej filozofii umysłu, a następnie poddać to rozumienie krytyce oraz zaproponować alternatywę dla niego.

Warto zacząć od wprowadzenia ważnej dystynkcji, która pozwoli doprecyzować cel tego rozdziału oraz wyznaczy ramy pojęciowe dla wywodu w nim prowadzonego. Chodzi o zaproponowane przez Daniela Dennetta (1995) odróżnienie *osobowego* i *subosobowego* poziomu wyjaśniania[2]. Jak stwierdza Dennett:

> Gdybyśmy powiedzieli, że osoba posiada doznanie bólu, umiejscawia je i skłonna jest do reagowania w pewien sposób, powiedzielibyśmy

[2] Rozróżnienie Dennetta rodzi pewne trudności interpretacyjne i często nie jest wykorzystywane w sposób do końca jasny i jednoznaczny (Drayson 2012). Samemu Dennettowi zarzucano, że w późniejszych pracach posługiwał się tą dystynkcją w sposób zasadniczo odbiegający od pierwotnej idei (chodzi o to, że w pewnym momencie badacz zaczął traktować treść stanów osobowych jako swoisty podzbiór treści intencjonalnej stanów subosobowych; por.: McDowell 1994; Hornsby 2000). Obecnie poprzestanę na przedstawieniu szkicowej, możliwie neutralnej rekonstrukcji tego rozróżnienia. Szczegółowa propozycja dotycząca tego, jak dokładnie należy rozumieć opozycję osobowe–subosobowe, zostanie przedstawiona w podrozdziale 5.2.

> już wszystko, co da się powiedzieć w ramach tego sposobu mówienia. Możemy domagać się dalszego wyjaśnienia, jak to się dzieje, że odsuwamy rękę od gorącego pieca, ale nie możemy domagać się dalszego wyjaśnienia w terminach „procesów mentalnych". [...] Jeżeli się na to decydujemy, musimy porzucić ten poziom wyjaśniania, poziom, w którym mówimy o ludziach, ich wrażeniach, ich działaniach, i zejść na subosobowy poziom mózgu i zdarzeń zachodzących w układzie nerwowym. Lecz gdy opuszczamy poziom osobowy, to zupełnie dosłownie opuszczamy również dziedzinę bólu [...], gdyż nasza alternatywna analiza nie jest wcale analizą bólu, ale czegoś zupełnie innego: ruchów ludzkiego ciała, czy też organizacji układu nerwowego (Dennett 1995: 106–107).

Powyższy cytat jest dla bieżących celów o tyle niefortunny, że jako przykład stanu osobowego jest w nim wskazane doznanie bólu, a nie postawa propozycjonalna. Aby odpowiednio ukierunkować te rozważania, oprócz wyjaśnień wymienionych w powyższym cytacie rozważmy także cztery inne:

> **(O1)** Jan pierwszy raz w życiu zagłosował na socjaldemokratów, ponieważ uznał, że wcześniej mylił się, pozytywnie oceniając neoliberalizm.
> **(O2)** Jan podniósł rękę na zebraniu spółdzielni mieszkaniowej, ponieważ pragnął zabrać głos w sprawie nowej strategii segregacji śmieci.
> **(S1)** Jan potrafi skutecznie chwytać przedmioty dzięki temu, że jego system motoryczny korzysta z emulatora ciała własnego (systemu mięśniowo-szkieletowego).
> **(S2)** Jan nie potrafi nawiązywać trwałych, opartych na przywiązaniu emocjonalnym relacji społecznych ze względu na doznane w dzieciństwie mechaniczne uszkodzenie ciała migdałowatego.

Wyjaśnienia (O1) i (O2) są sformułowane na poziomie osobowym. Co je charakteryzuje? Po pierwsze, wyjaśnienia te stanowią odpowiedź na pytanie o to, *dlaczego* Jan zachował się w określony sposób. Po drugie, ich eksplanandami są *działania*, których autorstwo przy-

pisujemy Janowi[3]. Po trzecie, eksplananse tych wyjaśnień odwołują się do stanów mentalnych, które mogą być w sposób poprawny i niemetaforyczny orzekane jedynie o *osobach* (podmiotach intencjonalnych). Przypisywanie neuronowi, populacji neuronów czy jakiemukolwiek wewnętrznemu narządowi Jana (lub komponentowi takiego narządu) przekonań na temat doktryn polityczno-ekonomicznych stanowiłoby błąd kategorialny. Neurony ani populacje neuronów nie mają przekonań, pragnień ani wątpliwości[4].

Należy zwrócić uwagę na związek między poziomem osobowym a psychologią potoczną. Podobnie jak posługiwanie się psychologią potoczną – formułowanie wyjaśnień na poziomie osobowym nie wymaga formalnej edukacji psychologicznej. Wyjaśnianie na poziomie osobowym – tak samo jak wyjaśnianie za pomocą psychologii potocznej – jest stosowane w toku codziennych interakcji społecznych. Kategorie, do jakich odwołują się wyjaśnienia na poziomie osobowym, to te same kategorie stanów mentalnych, które zawiera psychologia potoczna. W szczególności wyjaśnienia osobowe – co widać na przedstawionych przykładach – mogą odwoływać się do postaw propozycjonalnych. Można zatem przyjąć, że osobowy poziom wyjaśniania to inaczej poziom wyjaśniania, na którym jest aplikowana aparatura pojęciowa psychologii potocznej.

Spójrzmy teraz na zdania (S1) i (S2). Możemy przyjąć, że (S1) i (S2) to swego rodzaju „eliptyczne" (pomijające ogrom neurofizjo-

[3] Zawężanie klasy eksplanandów wyjaśnień osobowych do działań może zostać uznane za zbyt restrykcyjne i ograniczone. Z psychologii potocznej korzystamy, nie tylko wyjaśniając, dlaczego ktoś postąpił w dany sposób, ale też wyjaśniając, dlaczego osoba ta posiada przekonania czy pragnienia o określonej treści. Innymi słowy, postawy propozycjonalne wyjaśniają nie tylko działania, ale także *inne postawy propozycjonalne*. Moją intencją nie jest negowanie tego oczywistego faktu. Zawężam się do mówienia o „działaniach" dla uproszczenia wywodu, a nie z ważnych powodów teoretycznych. Można przyjąć, że używana tu kategoria „działań" obejmuje także osobowe procesy inferencyjne, polegające na przyjmowaniu postaw propozycjonalnych na podstawie innych posiadanych postaw propozycjonalnych.

[4] Nawiasem mówiąc, analogicznie jest z przynajmniej niektórymi *zdolnościami* poznawczymi: to nie kora wzrokowa w mózgu Jana, lecz Jan wzrokowo percypuje przedmioty; to nie skrzyżowanie skroniowo-ciemieniowe w jego mózgu, lecz Jan przypisuje innym ludziom stany mentalne.

logicznych czy neuroobliczeniowych szczegółów) wersje wyjaśnień, jakie są formułowane na poziomie *subosobowym*. Co charakteryzuje takie wyjaśnienia? Po pierwsze, choć skrótowa forma (S1) i (S2) nieco skrywa ten fakt, to powinniśmy je rozumieć jako wyjaśnienia tego, jak to się dzieje, że, odpowiednio, Jan posiada zdolność skutecznego chwytania przedmiotów oraz nie potrafi nawiązywać relacji emocjonalnych z innymi ludźmi. Na przykład gdybyśmy rozwinęli (S1), pokazalibyśmy, jak wykorzystanie emulatora ciała pozwala systemowi motorycznemu Jana generować ruchy pozwalające na skuteczne chwytanie obiektów w otoczeniu. Po drugie, wyjaśnienia te wskazują wewnętrzne warunki umożliwiające (*enabling conditions*) posiadanie określonych *zdolności* (lub dysfunkcji) poznawczych czy behawioralnych (McDowell 1994; Hurley 2008). Na przykład (S1) nie wyjaśnia, dlaczego w danych okolicznościach Jan sięgnął po konkretną książkę (działanie), lecz wyjaśnia, jak to się dzieje, że Jan w ogóle potrafi chwytać przedmioty (zdolność). Po trzecie, eksplananse, na które powołują się (S1) i (S2), to nie tyle stany mentalne *Jana*, ile raczej komponenty lub stany komponentów jego *ośrodkowego układu nerwowego*. Na przykład zgodnie z (S2) to nie fakty o Janie jako osobie, lecz fakty o jego ciele migdałowatym wyjaśniają niezdolność do nawiązywania związków emocjonalnych z innymi. Wyjaśnienia na poziomie subosobowym korzystają z kategorii neurofizjologicznych, obliczeniowych czy neuroobliczeniowych. Odwołują się one do wewnętrznej, strukturalno-funkcjonalnej architektury systemów poznawczych.

Filozofowie posługujący się dystynkcją osobowe–subosobowe twierdzą często, iż oba poziomy wyjaśniania są przynajmniej w jakimś stopniu wzajemnie *autonomiczne* (por.: McDowell 1994; Dennett 1995; Hornsby 2000). W szczególności wyjaśnienia formułowane na poziomie osobowym są odrębne od wyjaśnień na poziomie subosobowym oraz niezależne w stosunku do nich. Można sformułować wyczerpujące wyjaśnienie pewnego działania w kategoriach z poziomu osobowego, całkowicie pomijając kategorie czy fakty subosobowe. Uwidacznia się to w przytoczonym wyżej cytacie z Dennetta. Kiedy wyjaśniamy odsunięcie ręki od palącego się płomienia, stwierdzając, że dana osoba doznała bólu, nasze wyja-

śnienie jest już kompletne na poziomie osobowym. Nie istnieje *w ramach tego poziomu* możliwość odpowiedzi na pytanie o to, dlaczego odczuwamy ból w określonych okolicznościach. Jedyna możliwa „odpowiedź" brzmi bowiem: po prostu tak jest. Wyjaśnienie takie powinno być więc uznane za kompletne i zrozumiałe nawet w sytuacji, gdy nie zostanie ono uzupełnione twierdzeniami dotyczącymi neurofizjologii bólu. Analogicznie, kiedy wyjaśniamy przekonanie pewnej osoby, próbując pokazać, jak wywnioskowała je ona z innych przekonań (przekonań, które stanowiły racje dla przyjęcia przekonania wyjaśnianego), takie wyjaśnienie może być uznane za wyczerpujące. Powołanie się na przyczynową rolę obliczeń zachodzących w korze czołowej tej osoby byłoby nieuprawnione. Jak to ujmuje Dennett, zmienilibyśmy w ten sposób „temat" i nie mielibyśmy już w naszym wyjaśnianiu do czynienia z myślącymi osobami.

Można teraz wykorzystać dystynkcję osobowe–subosobowe, aby dookreślić cel tego rozdziału. Zwróćmy przede wszystkim uwagę na zbieżność między wspomnianą opozycją a Sellarsowskim odróżnieniem dwóch obrazów świata. Można powiedzieć, że linia oddzielająca to, co się manifestuje, od tego, co naukowe, jest zarazem linią oddzielającą to, co osobowe, od tego, co subosobowe[5]. Manifestującemu się obrazowi umysłu odpowiada poziom osobowy, a zatem poziom psychologii potocznej. Jest to obraz wykorzystywany przez ludzi w toku ich interakcji z innymi ludźmi, interakcji, które przynajmniej w jakimś stopniu opierają się na atrybuowaniu postaw propozycjonalnych. Z kolei naukowy obraz zawiera się w teoriach i modelach formułowanych na poziomie subosobowym. Jest to rysowany przez nauki kognitywne wizja wewnętrznej, mechanistycznej architektury systemów poznawczych. Zamiast pytać o relację między dwoma Sellarsowskimi obrazami świata, w dalszej części tego rozdziału będę pytać raczej o relację między osobowym a subosobowym poziomem wyjaśniania. Traktuję co prawda te zagadnienia jako w przybliżeniu tożsame, jednak sformułowanie odwołujące

[5] Zawężamy się tu rzecz jasna tylko do tego, jak w ramach obu obrazów jest postrzegana natura umysłów czy poznania.

się do opozycji osobowe–subosobowe wydaje się bardziej precyzyjne i określone, a przez to – analitycznie owocniejsze.

Mówiąc, że zaproponowana w rozdziale 4 teoria wyjaśniania reprezentacyjnego sytuuje się w obrębie obrazu naukowego, miałem na myśli przede wszystkim to, iż stanowi ona koncepcję wyjaśniania *subosobowego*. Postulowane tu modele mentalne są reprezentacjami o charakterze subosobowym i pełnią one funkcję w subosobowych wyjaśnieniach. Dlaczego? Po pierwsze, MKM nie mają wyjaśniać tego, dlaczego ludzie podejmują określone działania, lecz to, *jak* systemy poznawcze realizują zadania, w których manifestują się posiadane przez nie *zdolności* (ewentualnie, jak dochodzi do określonych dysfunkcji). Po drugie, zarówno modele wewnętrzne (nośniki modeli), jak i ich konsumenci to pewne struktury o charakterze subosobowym, to znaczy *komponenty wewnętrznych mechanizmów* poznawczych. Nie można powiedzieć, że *osoba* jest (jakkolwiek by to rozumieć) „podmiotem" modelu mentalnego albo że posługuje się tym modelem w tym samym sensie, jak może ona posługiwać się zewnętrzną mapą. Modelami wewnętrznymi w przyjętym tu znaczeniu „posługują" się konsumenci reprezentacji, a zatem subosobowe komponenty wewnętrznych mechanizmów poznawczych. Po trzecie wreszcie, przeznaczeniem bronionej tu teorii mechanizmów reprezentacyjnych nie jest „naturalizacja" postaw propozycjonalnych. Dopuszczam możliwość, że odkrywane przez kognitywistów subosobowe modele wewnętrzne będą charakteryzować się własnościami funkcjonalnymi oraz intencjonalnymi, które zasadniczo różnią się od własności przypisywanych stanom z poziomu osobowego (postawom propozycjonalnym). Na przykład jeśli modele są emulatorami, to w niczym nie przypominają one przekonań czy pragnień. Nie stanowi to dla teorii MKM problemu, ponieważ nie dąży ona do uzgodnienia naukowej wizji systemu poznawczego z psychologią potoczną. Jest ona w zamierzeniu teorią wyjaśnień subosobowych, a nie próbą naturalizacji intencjonalności.

Dla niektórych zwolenników projektu naturalizacji intencjonalności – albo dla filozofów nieodróżniających go w praktyce od problemu eksplanacyjnego statusu reprezentacji w kognitywistyce – teoria mechanizmów wykorzystujących wewnętrzne modele

może być w pewien sposób rozczarowująca. Nie rozwija ona Sellarsowskiego projektu uzgodnienia dwóch obrazów w ramach jednolitej wizji. Skupia się na poziomie subosobowym, pomijając kwestię możliwości pogodzenia go z osobowym. Na czym jednak takie pogodzenie miałoby w ogóle polegać? Czego dokładnie oczekujemy, kiedy mówimy, że to, co osobowe (manifestujące się), powinno zostać „uzgodnione" czy „pogodzone" z tym, co subosobowe (naukowe)? Czego *powinniśmy* oczekiwać? Czy taka koncyliacja wymaga tego, aby problem eksplanacyjnego statusu reprezentacji w kognitywistyce został rozwiązany w jakiś ściśle określony sposób? Pytania te wyznaczają obszar tematyczny obecnego rozdziału. Jak zobaczymy, spora część współczesnej analitycznej filozofii umysłu opiera się na określonym założeniu dotyczącym międzypoziomowych zależności zachodzących między tym, co osobowe, a tym, co subosobowe. Chcę to założenie zidentyfikować i opisać (w sekcji 5.1.2), a także wykazać, iż nie ma ono tak mocnych podstaw, jak się wydaje (w sekcji 5.1.3). Celem tego rozdziału jest też zaproponowanie (w podrozdziałach 5.2 i 5.3) alternatywnego, opartego na mechanicyzmie sposobu rozumienia poziomów osobowego i subosobowego oraz ich wzajemnych związków.

5.1.2. Naturalizowanie intencjonalności a założenie o korespondencji osobowe–subosobowe

W rozdziale 1 problem naturalizacji intencjonalności został odróżniony od zagadnienia eksplanacyjnej użyteczności reprezentacji w kognitywistyce. Przypomnijmy sobie to rozróżnienie (por. tabela 1). Celem projektu naturalizacji intencjonalności jest dostarczenie koncepcji przedstawiającej *stricte* naturalne warunki, których spełnienie wystarcza do tego, by dany stan mógł zostać zidentyfikowany z postawą propozycjonalną o określonej treści (domyślnie te warunki miałyby spełniać wewnętrzny stan lub struktura systemu poznawczego). Problem ten skupia się zatem na osobowych stanach intencjonalnych, czyli postawach propozycjonalnych. Projekt naturalizacji opiera się także na traktowaniu reprezentacji mentalnych (postaw propozycjonalnych) jako eksplanandum, czyli przedmio-

tu szeroko pojętego wyjaśniania. Owe fakty odróżniają *prima facie* naturalizację intencjonalności od problemu eksplanacyjnej użyteczności reprezentacji dla kognitywistyki. Ten ostatni problem stanowi wewnętrzne zagadnienie nauk o poznaniu, usytuowane w ramach obrazu naukowego i skupia się na pytaniu o to, czy pojęcie reprezentacji mentalnych jest kognitywistom przydatne do wyjaśniania interesujących ich zjawisk. Przedmiotem zainteresowania są reprezentacje o charakterze subosobowym, które nie muszą przypominać intencjonalnie i/lub funkcjonalnie stanów wyróżnianych w ramach psychologii potocznej. Wreszcie problem ten traktuje reprezentacje przede wszystkim jako eksplananse, na które powołują się (mechanistyczne) wyjaśnienia zjawisk formułowane w ramach kognitywistyki.

Wszystkie te różnice sprawiają, że naturalizacja intencjonalności i kwestia statusu eksplanacyjnego reprezentacji w kognitywistyce powinny być traktowane jako zagadnienia oddzielne i przynajmniej potencjalnie niezależne. Samo ich odróżnienie nie przesądza jednak o tym, czy (oraz jak) są one powiązane. Jak już zobaczyliśmy w rozdziale 1 (podrozdział 1.2), w literaturze filozoficznej ostatnich dekad dość powszechne jest przekonanie o ścisłym związku tych problemów. Wielu filozofów zaangażowanych w projekt naturalizacji przyjmuje (bardziej lub mniej *explicite*), że jego sukces pozostaje ściśle uzależniony od tego, czy możliwa jest naukowa „rehabilitacja" wyjaśnień odwołujących się do intencjonalnych kategorii psychologii potocznej. W takim ujęciu wyjaśnienia działań odwołujące się do przekonań i pragnień powinny okazać się, mówiąc bardzo szeroko, zgodne czy ciągłe z wyjaśnieniami naukowymi. Jeśli taka naukowa rehabilitacja psychologii potocznej się nie powiedzie, może to prowadzić do konieczności eliminacji postaw potocznych z naukowego obrazu świata. Innymi słowy, sukces projektu naturalizacji zależy od tego, czy postawy propozycjonalne okażą się naukowo wartościowym rodzajem reprezentacji, na które mogą powoływać się *kognitywistyczne* wyjaśnienia ludzkich działań. Na czym opiera się takie ujęcie relacji między naturalizacją intencjonalności a problemem statusu eksplanacyjnego reprezentacji w kognitywistyce? Jakie filozoficzne założenie stoi u jego podstaw?

Podejmując kwestię związku między kognitywistyką a naturalizacją intencjonalności, Peter Godfrey-Smith (2004) odróżnił dwa rodzaje faktów: (1) fakty architekturalne (*wiring-and-connection facts*), dotyczące wewnętrznej organizacji złożonych behawioralnie organizmów oraz związku tej organizacji ze środowiskiem zewnętrznym; (2) fakty dotyczące interpretacji, czyli praktyk polegających na interpretowaniu innych przez przypisywanie im stanów posiadających treść (postaw propozycjonalnych). Według Godfreya-Smitha dla wielu współczesnych filozofów (1) i (2) są ze sobą ściśle powiązane – przyjmują oni bowiem, że praktyka interpretacji w kategoriach intencjonalnych niesie określone *zobowiązania architekturalne*. Sądzą przy tym, że (a) przypisanie komuś postawy propozycjonalnej stanowi próbę powiedzenia czegoś na temat wewnętrznych, architekturalnych faktów dotyczących systemu poznawczego tej osoby; (b) prawomocność praktyki interpretacji w kategoriach psychologii potocznej zależy od tego, czy towarzyszące tej praktyce zobowiązania architekturalne są poprawne. Praktyka interpretacji w potocznych kategoriach intencjonalnych jest więc nieuprawniona, jeśli na poziomie wewnętrznej architektury poznania nie ma nic, co odpowiadałoby (funkcjonalnie/intencjonalnie) stanom wyróżnianym przez psychologię potoczną. Moglibyśmy zatem powiedzieć, że z takiego punktu widzenia celem projektu naturalizacji intencjonalności jest sformułowanie odpowiedzi na pytanie o to, jakie powinny zachodzić fakty architekturalne, aby ludzkie praktyki interpretacyjne mogły zostać uprawomocnione.

Spróbujmy przeformułować diagnozę Godfreya-Smitha za pomocą aparatury pojęciowej wprowadzonej w poprzedniej sekcji. Proponuję uznać, że wyróżnione przez tego autora fakty architekturalne to inaczej fakty, na które powołują się wyjaśnienia z poziomu subosobowego. Wyjaśnienia te odwołują się wszakże do wewnętrznej, mechanistycznej architektury systemu poznawczego. Z kolei fakty dotyczące interpretacji dotyczą praktyk opisywania i wyjaśniania działań na poziomie osobowym, z użyciem potocznych kategorii psychologicznych. Jeśli przystaniemy na takie odczytanie dystynkcji Godfreya-Smitha, to filozoficzne założenie, o którym mówi ten autor, możemy określić jako *założenie o międzypoziomowej kore-*

spondencji między poziomem osobowym a subosobowym (w skrócie: ZKOS lub „założenie korespondencji"). Sądzę, że ZKOS ma zarówno aspekt epistemiczny, dotyczący poziomów wyjaśniania, jak i metafizyczny, dotyczący natury stanów i struktur, na które powołują się wyjaśnienia na obu poziomach. Oto jak rozumiem założenie korespondencji w obu tych aspektach.

Metafizyczne sformułowanie ZKOS

Być podmiotem postawy propozycjonalnej, to mieć „w głowie" stan subosobowy o określonych własnościach. Dowolna osoba jest podmiote przekonania czy pragnienia wtedy, gdy w subosobowej architekturze jej systemu poznawczego można wyróżnić komponent lub stan komponentu, który moglibyśmy zidentyfikować z tym przekonaniem czy pragnieniem[6]. Istnienie przekonań, pragnień i innych

[6] Ktoś mógłby zwrócić uwagę, że dla wielu współczesnych filozofów postawy propozycjonalne są jedynie egzemplarycznie identyczne z subosobowymi stanami czy strukturami mózgu. Każdy egzemplarz postawy propozycjonalnej jest identyczny z jakimś egzemplarycznym stanem neuronalnym, jednak to, co metafizycznie konstytutywne czy istotne dla bycia postawą propozycjonalną, stanowi pewną własność wyższego rzędu (na przykład obliczeniową). Nie jest to własność neuronalna, choć może być ona (wielorako) realizowana przez własności neuronalne. Z takiego punktu widzenia neuronauka dostarcza tylko stosunkowo nieistotnych szczegółów dotyczących implementacji postaw propozycjonalnych, ale nie mówi nic ciekawego o ich *naturze*. Czy w takiej sytuacji również mamy do czynienia z ZKOS? Czy jest w ogóle użyteczne przypisywanie ZKOS koncepcjom, które uznają tezę o zaledwie egzemplarycznej, a zatem w pewnym sensie teoretycznie czy eksplanacyjnie nieciekawej identyczności między stanami osobowymi a neuronalnymi?

Powyższa wątpliwość opiera się na zbyt wąskim postrzeganiu stanów subosobowych. Przede wszystkim wszelkie stany obliczeniowe kwalifikują się jako subosobowe. Jak najbardziej subosobowe są zatem Fodorowskie zdania wyrażone w języku myśli. W takim ujęciu nawet dla Fodora występuje teoretycznie znacząca identyczność stanów subosobowych z osobowymi na poziomie *typicznym* czy *rodzajowym*, a nie jedynie egzemplarycznym (przekonanie, że *p*, pojęte jako rodzaj ma za swój odpowiednik pewien rodzaj stanu obliczeniowego). Co więcej, stany czy własności obliczeniowe to zawsze stany/własności *komponentów* systemu poznawczego (por. przedstawione w sekcji 2.3.2 omówienie mechanistycznego spojrzenia na wyjaśnienia obliczeniowe oraz zagadnienie wielorakiej realizacji). W przypadku biologicznych systemów poznawczych będą to komponenty ośrodkowego układu nerwowego (zakładając

postaw propozycjonalnych zależy od tego, czy odpowiadają im jakieś stany z poziomu subosobowego (neuronalne czy neurooboliczeniowe). Ujmując tę kwestię bardziej technicznie:

> Dla dowolnego podmiotu *S* i dowolnej postawy propozycjonalnej *P*, podmiot *S* posiada postawę propozycjonalną *P* wtedy i tylko wtedy, gdy wewnątrz *S* istnieje pewien subosobowy stan czy struktura *N*, która ma odpowiednie własności intencjonalne/funkcjonalne. (Można powiedzieć, że *N* stanowi subosobowy odpowiednik *P*).

Epistemiczne sformułowanie ZKOS

Wyjaśnienia wykorzystujące psychologię potoczną jedynie pozornie różnią się od tych formułowanych w kognitywistyce. Wyjaśnienia odwołujące się do przekonań i pragnień niosą określone zobo-

---

rzecz jasna, że systemy te rzeczywiście działają obliczeniowo). Na gruncie założeń przyjmowanych w tej pracy nie można zatem mocno oddzielać wyjaśnień obliczeniowych od neuronaukowych. Co więcej, na bieżące potrzeby należałoby odróżnić neuronaukę jako dyscyplinę *stricte* biologiczną od neuronauki jako dyscypliny kognitywnej (por. Gold, Stoljar 1999). Ta pierwsza zajmuje się jedynie biologicznymi oraz biochemicznymi aspektami działania mózgu. Ta druga traktuje jednak mózg jako system zajmujący się realizacją funkcji *poznawczych*. Dyscyplina ta może stosować do badania oraz opisu mózgu między innymi kategorie obliczeniowe. Na jej gruncie rodzaje czy kategorie stanów neuronalnych (subosobowych) mogą być wyróżniane na podstawie ról funkcjonalnych (także obliczeniowych), a nawet własności intencjonalnych. A zatem ten typ neuronauki przypisuje strukturom subosobowym własności, które potencjalnie mogą „upodabniać" je do postaw propozycjonalnych. Mowa zatem o strukturach, wobec których możemy całkiem sensownie zapytać, *dlaczego* czy też *na podstawie jakich własności* funkcjonalnych/intencjonalnych struktury te mają być identyczne z postawami propozycjonalnymi. Kiedy poruszamy kwestię relacji osobowe–subosobowe w kontekście tak pojmowanej neuronauki, nie pytamy już o „banalne" szczegóły implementacyjne, a zajmujemy się całkiem treściwym zagadnieniem, czy *mózg działa w sposób* (w jakimś sensie czy zakresie) *„zgodny" z wizją umysłu zawartą w psychologii potocznej*. Pytamy więc między innymi o to, czy „rodzaje naturalne" psychologii potocznej mogą być utożsamione z jakimiś „rodzajami naturalnymi" kognitywistyki (w tym neuronauki poznawczej). Założenie korespondencji stanowi zatem zawsze zupełnie nietrywialną presupozycję teoretyczną.

wiązania dotyczące wewnętrznej, subosobowej organizacji systemu poznawczego. Wyjaśnienia osobowe są swego rodzaju „zawoalowanymi" wyjaśnieniami subosobowymi. Aby dowolne wyjaśnienie na poziomie osobowym mogło zostać uznane za poprawne, musi mu odpowiadać pewne (poprawne) wyjaśnienie sformułowane na poziomie subosobowym. Mówiąc bardziej technicznie:

> Dla dowolnego osobowego wyjaśnienia *O*, które wyjaśnia eksplanandum *E* za pomocą postawy propozycjonalnej *P*, wyjaśnienie to jest poprawne (prawdziwe) wtedy i tylko wtedy, gdy poprawne (prawdziwe) jest subosobowe wyjaśnienie *SO*, które wyjaśnia *E* za pomocą stanu czy struktury *N*, stanowiącej subosobowy odpowiednik *P*.

Sądzę, że charakteryzowanie założenia korespondencji jako doktryny zarówno metafizycznej, jak i dotyczącej wyjaśniania jest jak najbardziej zbieżne z tym, jak założenie to *de facto* funkcjonuje we współczesnej filozofii i kognitywistyce. Projekt naturalizacji intencjonalności bywa całkiem *explicite* definiowany jako zarazem zmierzający: (1) do znalezienia dla postaw propozycjonalnych „miejsca" w świecie fizycznym oraz (2) do pokazania, jak wyjaśnienia odwołujące się do postaw propozycjonalnych mogą zostać „zrehabilitowane" przez naukę (por.: Fodor 1987; Dretske 1988). Co więcej, zgodnie z niemal konsensualnie utrzymywanym współcześnie poglądem wyjaśnienia działań za pomocą pojęć psychologii potocznej mają charakter przyczynowy (por.: Davidson 1963; Fodor 1987; Dretske 1988). Wyjaśniając czyjeś działanie za pomocą przekonania o określonej treści, wskazujemy przyczynę tego działania. Jakiego rodzaju jest to przyczyna? To subtelne i złożone zagadnienie, jednak możemy w pewnym uproszczeniu powiedzieć, że dla wielu autorów teza o przyczynowej naturze wyjaśnień na poziomie osobowym będzie możliwa do utrzymania tylko wtedy, gdy uznamy, iż postawy propozycjonalne są identyczne z subosobowymi strukturami mózgu (por. szczegółowe i krytyczne omówienie takiego podejścia w: Baker 1995: 3–32, 93–150). Abstrahując teraz od poprawności takiego podejścia (zagadnienie to zostanie poruszone szerzej w sekcji 5.3.3),

zwróćmy tylko uwagę, iż ściśle wiąże ono określone spojrzenie na *wyjaśnianie* na poziomie osobowym (jako na wyjaśnianie przyczynowe) ze spojrzeniem na *naturę* stanów osobowych (jako identycznych ze stanami subosobowymi).

We współczesnej filozofii umysłu założenie korespondencji pojawia się w dwóch zasadniczych kontekstach. W każdym z nich ma ono inne zastosowanie czy też służy odmiennemu celowi. Pierwsze z zastosowań ma charakter *konstruktywny*. Chodzi o rolę, jaką spełnia ZKOS w projekcie naturalizacji intencjonalności. Filozofowie zaangażowani w ten projekt chcą pokazać, jak jest możliwe istnienie intencjonalnych stanów osobowych w świecie fizycznym, opisywanym i wyjaśnianym przez nauki szczegółowe. ZKOS przejawia się nie tyle w celu projektu naturalizacji, ile raczej w powszechnie akceptowanym sposobie pojmowania środka, za pomocą którego należy to zamierzenie zrealizować. Środkiem tym jest, przypomnijmy, wyrażenie – w kategoriach naturalistycznych i niepresuponujących intencjonalności – warunków wystarczających do bycia postawą propozycjonalną. Zapytajmy zatem: co wedle zwolenników projektu naturalizacji miałoby potencjalnie spełniać takie warunki, jakiekolwiek by one nie były? Jakie rodzaje naturalnych struktur w świecie są „kandydatami" do spełnienia tych warunków i bycia przez to „znaturalizowanymi" postawami propozycjonalnymi? Gdyby ktoś pragnął odpowiedzieć na te pytania na tyle ogólnie, by odpowiedź obejmowała wszystkie znaczące teorie rozwijane w ramach projektu naturalizacji na przestrzeni ostatnich dziesięcioleci, musiałby powiedzieć: kandydatami takimi są stany czy struktury subosobowe. Rzecz jasna, nie istnieje konsensus co do tego, o jakie stany subosobowe chodzi. Dla Jerry'ego Fodora (1975, 1987, 2001) są to zdania wyrażone w wewnętrznym języku myśli, które znajdują się w odpowiedniej „skrzynce" funkcjonalnej (na przykład przekonaniowej). Dla Freda Dretskego (1981, 1988) są to funkcjonalne stany informacyjne organizmu. Propozycja Ruth Millikan (1984, 2002) jest pod wieloma względami zbliżona do teorii Dretskego, lecz autorka ta rozwija wątek teleologiczny i nadaje szczególne znaczenie doborowi natural-

nemu[7]. Dla zwolenników semantyki ról funkcjonalnych – takich jak Ned Block (1986) czy Gilbert Harman (1987) – postawy propozycjonalne to wewnętrzne struktury odgrywające w systemie poznawczym odpowiednie role przyczynowe. Mimo wszystkich różnic między wymienionymi teoriami – istnieje wspólne im założenie, którym jest właśnie ZKOS. Wszystkie te koncepcje wspierają się na ZKOS jako fundamentalnej – moglibyśmy powiedzieć: paradygmatycznej – presupozycji. Filozofowie zaangażowani w projekt naturalizacji intencjonalnością sądzą, że przypisywane ludziom przekonania i pragnienia – o ile w ogóle ktokolwiek kiedykolwiek rzeczywiście miał jakieś przekonanie lub czegoś pragnął – muszą okazać się w istocie identyczne z jakimiś stanami obliczeniowymi albo biologicznymi wyróżnianymi na poziomie subosobowym. Czym innym mogą być wszakże przekonania, jeśli nie stanami biologicznymi albo obliczeniowymi organizmu? Założenie korespondencji przyjmuje się w projekcie naturalizacji intencjonalności za tak oczywiste, że na ogół nie jest nawet *explicite* wyrażane.

---

[7] Powstaje tu wątpliwość, czy takie przedstawienie teorii Dretskego i Millikan nie pozostaje zbyt wąskie. Czy przedmiotem zainteresowania tych autorów nie jest naturalizacja własności intencjonalnych w ogóle? Teorie tych autorów dotyczą przecież nie tylko przekonań i pragnień, ale też stosunkowo prostych reprezentacji (a w każdym razie struktur postulowanych jako reprezentacje), na przykład takich, jakimi mają posługiwać się termostaty albo układy wzrokowe żab. Jest to celna obserwacja, jednak co najwyżej dodatkowo potwierdza ona obecność założenia korespondencji w teoriach Dretskego i Millikan. Autorzy ci chcą wskazać naturalne warunki wystarczające do bycia stanem intencjonalnym. W ten sposób chcą oni jednak znaturalizować także postawy propozycjonalne. Nie jest tak, że (naturalne) własności determinujące treść postaw propozycjonalnych są różne od własności, które determinują treść prostych detektorów w mózgu żaby. Między prostymi reprezentacjami a postawami propozycjonalnymi zachodzą teoretycznie ważne różnice, jednak nie pojawia się między nimi jakaś jakościowa przepaść. Innymi słowy, Dretske i Millikan nie traktują stanów osobowych i subosobowych jako zasadniczo czy fundamentalnie różnych. Z ich perspektywy postawy propozycjonalne są wewnętrznymi, funkcjonalnymi stanami informacyjnymi organizmów, tak samo jak prostsze reprezentacje w rodzaju żabich detektorów. Zarówno reprezentacje w korze wzrokowej żab, jak i postawy propozycjonalne u ludzi są dla Dretskego i Millikan zlokalizowanymi „w głowie", posiadającymi treść, przyczynowo aktywnymi, subosobowymi komponentami czy stanami komponentów systemu poznawczego.

Drugi rodzaj zastosowania ZKOS we współczesnej filozofii umysłu ma charakter *destruktywny*. Chodzi tu o funkcję, jaką założenie korespondencji pełni dla filozofów argumentujących za koniecznością *eliminacji* kategorii psychologii potocznej z naukowego obrazu świata. Mowa konkretnie o formie eliminatywizmu, którą można określić jako eliminatywizm architekturalny[8]. Kwalifikuję to stanowisko w taki sposób ze względu na kluczową rolę, jaką w argumentacji na jego rzecz odgrywają fakty dotyczące wewnętrznej, subosobowej architektury systemu poznawczego. Eliminatywizm architekturalny cechuje się tym, że na jego rzecz jest zawsze przytaczany argument stworzony na podstawie następującego schematu:

1. Dla bycia postawą propozycjonalną jest konieczne posiadanie własności *W*.
2. Subosobowe stany i struktury odkrywane przez kognitywistów nie mają własności *W*.
3. Zatem subosobowe stany i struktury nie mogą być identyczne z postawami propozycjonalnymi.

[8] Eliminatywizm architekturalny powinien zostać odróżniony od stanowiska, które moglibyśmy nazwać „eliminatywizmem teoretycznym". Ta druga pozycja cechuje się tym, iż (1) opiera się na założeniu, że psychologia potoczna stanowi protonaukową teorię; (2) wskazuje na zasadnicze ograniczenia pojęciowe, predykcyjne czy eksplanacyjne tak rozumianej psychologii potocznej (na przykład na niemożność wyjaśnienia na jej gruncie natury i funkcji snów); (3) postuluje na tej podstawie, że psychologia potoczna (wraz z zakładaną przez nią taksonomią stanów mentalnych) powinna zostać odrzucona i zastąpiona teorią lepszą, pochodzącą z kognitywistyki (por.: Churchland 1981; Stich 1983). Czy także ten rodzaj eliminatywizmu opiera się na ZKOS? Choć nie jest to natychmiastowo widoczne, wydaje się, że można na to pytanie odpowiedzieć twierdząco. Traktowanie psychologii potocznej jako protonaukowej teorii oznacza między innymi uznanie, że wyjaśnienia na poziomie osobowym i subosobowym spełniają podobne funkcje epistemiczne. To znaczy, iż posługiwanie się psychologią potoczną to stosowanie teorii dotyczącej przyczynowych źródeł ludzkich działań. W domyśle chodzi tu jednak o przyczynowe źródła występujące na poziomie subosobowym. Psychologia potoczna to zatem zdroworozsądkowa, protonaukowa koncepcja tego, jak działa wewnętrzna, subosobowa architektura systemu poznawczego. Jeśli taka interpretacja tego stanowiska jest poprawna, to eliminatywizm teoretyczny (tak samo jak eliminatywizm architekturalny) opiera się na ZKOS, a konkretnie na ZKOS w epistemicznym sformułowaniu.

4. Zatem postawy propozycjonalne nie istnieją[9].

Zauważmy, że przejście między 1–3 a 4 jest entymematyczne. Jak brzmi ukryta przesłanka? Okazuje się nią właśnie ZKOS, a precyzyjniej – ZKOS w metafizycznym sformułowaniu. W eliminatywizmie architekturalnym przejście od przesłanek do wniosku jest uprawnione tylko wtedy, gdy przyjmujemy, że istnienie postaw propozycjonalnych zależy od tego, czy wewnętrzna, subosobowa architektura systemu poznawczego działa (przynajmniej częściowo) na podstawie struktur mających własności tych postaw (stanowiących ich subosobowe odpowiedniki). Jeżeli zatem struktury subosobowe nie posiadają lub nie mogą posiadać jakiejś własności istotnej czy koniecznej dla bycia postawą propozycjonalną, to okazuje się, że nie możemy przyjmować istnienia postaw propozycjonalnych.

Za klarowny przykład eliminatywizmu architekturalnego można uznać stanowisko Williama Ramseya, Stephena Sticha i Josepha Garona (1990), które zostało omówione w rozdziale 1 (podrozdział 1.2). Argument przedstawiony przez tych autorów podpada pod powyższy schemat i można go zrekonstruować w następujący sposób:

1. Postawy propozycjonalne cechują się z konieczności modularnością propozycjonalną[10].
2. W sieciach koneksyjnych nie występują stany czy struktury cechujące się modularnością propozycjonalną.
3. Subosobowa architektura systemu poznawczego przypomina sieć koneksyjną.

---

[9] Alternatywnie – wniosek może być bardziej ostrożny i stwierdzać jedynie, że przyjmowanie istnienia postaw propozycjonalnych jest niezgodne z naukowym obrazem świata. Dla wielu współczesnych naturalistycznie nastawionych filozofów te konkluzje są jednak równoważne (por. Baker 1995).

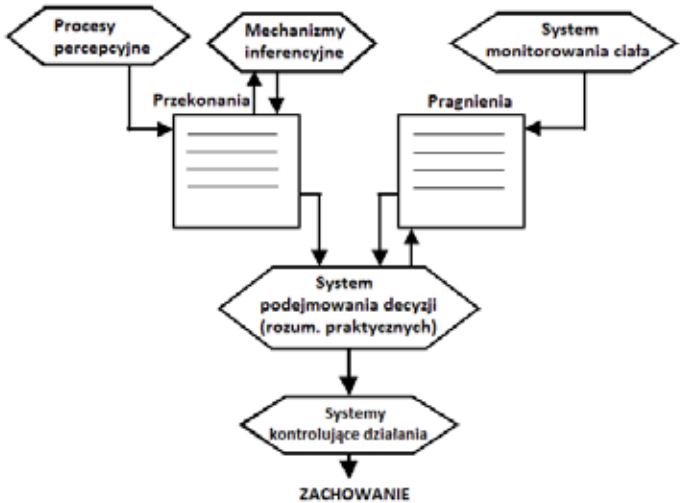
[10] Przypomnijmy, że modularność propozycjonalna polega na tym, iż (1) postawy propozycjonalne mogą być indywidualnie nabywane i odrzucane; (2) zawsze istnieje definitywna odpowiedź na pytanie o to, która z postaw propozycjonalnych posiadanych przez daną osobę stanowiła rzeczywistą przyczynę zachowania tej osoby.

4. Zatem subosobowe stany i struktury systemu poznawczego nie mogą być identyczne z postawami propozycjonalnymi.
5. Zatem postawy propozycjonalne nie istnieją.

Powyższy argument uzupełnia przedstawiony wcześniej schemat w dwóch miejscach. Po pierwsze, jako istotna własność osobowych stanów intencjonalnych zostaje w nim wymieniona modularność propozycjonalna. Po drugie, zawiera on twierdzenie, że subosobowa architektura systemu poznawczego przypomina budowę sieci koneksyjnej. Jednak w tym argumencie inferencyjne przejście od przesłanek do wniosku nadal wymaga założenia, że postawy propozycjonalne są identyczne z określonego rodzaju stanami czy strukturami subosobowymi. Można więc powiedzieć, że założenie korespondencji stanowi entymemat argumentu Ramseya, Sticha i Garona[11].

ZKOS stanowi zatem założenie wspólne filozofom zaangażowanym w naturalizację intencjonalności oraz eliminatywistom. Pomimo że ich cele są tak zasadniczo różne, oba te filozoficzne projekty wspierają się na założeniu korespondencji jako ważnej presupozycji. Na koniec zauważmy jednak, że wpływ ZKOS nie kończy się na samej filozofii. Założenie to „przedostaje" się także do teorii *stricte* kognitywistycznych. Przyjmuje w nich ono postać określonej wizji systemu poznawczego – każącej postrzegać go jako swoistą „skrzynkę" na postawy propozycjonalne. Teorie oparte na tej wizji wyjaśniają zjawiska przez odwołanie do subosobowej architektury, którą pojmują tak, jak gdyby realizowała ona coś w rodzaju wewnętrznej, „mechanicznej" psychologii potocznej.

[11] Co ciekawe, w pracy, która jest tu często przywoływana, Ramsey (2007: 38–66) argumentuje, że postawy propozycjonalne są „niekompatybilne" nie tylko z koneksjonizmem, ale także z GOFAI. Próbuje on zatem pokazać, że – wbrew często utrzymywanym opiniom – postawy propozycjonalne nie mogą być identyczne nawet ze strukturami postulowanymi w ramach klasycznej, symbolicznej wizji obliczeniowej architektury systemu poznawczego. Struktury postulowane w ramach GOFAI nie realizują bowiem w systemie „zadań" charakterystycznych dla przekonań i pragnień. Argumentacja Ramseya oraz jej związek z ZKOS zostały krytycznie omówione w: Gładziejewski 2012.



Rysunek 9. Stephena Sticha i Schauna Nicholsa szkic funkcjonalny procesu podejmowania decyzji. Źródło: Stich, Nichols 1992: 39

Weźmy pod uwagę konkretne przykłady takiej strategii. Spójrzmy chociażby na rysunek 9. Przedstawiony na nim diagram został zaczerpnięty z artykułu Stephena Sticha i Shauna Nicholsa (1992). Nie chcę tu rekonstruować tez ani argumentów w nim przedstawionych. Chodzi raczej o zwrócenie uwagi na zawarty tam ogólny sposób myślenia o wewnętrznej architekturze poznania. Diagram przedstawiony na rysunku 9 stanowi szkic funkcjonalny procesu podejmowania decyzji praktycznych. Możemy przyjąć, że szkic ten ma przedstawiać w ogólnym zarysie propozycję Sticha i Nicholsa dotyczącą organizacji procesów obliczeniowych stojących u podstaw podejmowania decyzji. Uderza w ich koncepcji fakt, że autorzy ci postrzegają subosobowe procesy obliczeniowe stojące u podstaw podejmowania decyzji przez dość ścisłą analogię do tego, jak proces ten jest opisywany na poziomie osobowym. Nasza potoczna psychologia dostarcza nam wizję, zgodnie z którą ludzie podejmujący decyzje praktyczne anga-

żują się w proces inferencyjnego „ważenia" czy uzgadniania posiadanych przekonań na temat świata oraz pragnień dotyczących tego, jak ten świat powinien wyglądać. Jeśli spojrzymy na diagram Sticha i Nicholsa, okazuje się, że z ich perspektywy te osobowe procesy inferencyjne są odzwierciedlone w obliczeniowych procesach subosobowych. Przekonaniom i pragnieniom odpowiadają dwa rodzaje stanów (neuro)obliczeniowych, które są przechowywane w funkcjonalnie odrębnych mentalnych „pojemnikach". Istnieje też jakiś rodzaj obliczeniowego modułu – niewykluczone, że zlokalizowany w jakimś dobrze wyróżnionym obszarze mózgu – który zajmuje się „przekładaniem" przekonań i pragnień na zachowania, a zatem tym, czym według psychologii potocznej zajmuje się podmiot intencjonalny. Opis w kategoriach przekonań, pragnień i procesów inferencyjnych okazuje się więc (bardziej lub mniej precyzyjnym) opisem tego, co dzieje się wewnątrz subosobowej maszynerii systemu poznawczego[12].

Zaprezentowany przykład nie jest rzadki czy odosobniony. Podobny sposób myślenia o systemie poznawczym uwidacznia się chociażby w niektórych koncepcjach dotyczących natury oraz źródeł stanów urojeniowych (*delusions*). Zwolennicy tak zwanej dwuczynnikowej teorii urojeń twierdzą, że stany urojeniowe mają dwa źródła: (1) zaburzenie subiektywnego doświadczenia, prowadzące do uformowania przekonania o urojeniowej treści (na przykład prze-

[12] Zdaję sobie sprawę, że diagramu Sticha i Nicholsa nie należy traktować zbyt dosłownie. Jest to zaledwie szkic funkcjonalny. Na przykład autorzy ci nie są zobowiązani do twierdzenia, że istnieje w ludzkim mózgu jakiś konkretny, wyraźnie rozgraniczony obszar zajmujący się tylko i wyłącznie nabywaniem, przechowywaniem i rewizją przekonań. Bez wątpienia funkcjonalny diagram mógłby zostać zrealizowany strukturalnie na wiele sposobów. Mimo to wydaje się jednak, że przynosi on z konieczności *pewne* nietrywialne zobowiązania dotyczące strukturalnej (a nie jedynie funkcjonalnej) organizacji systemu poznawczego (por. sekcja 3.3.2). Na przykład nawet jeśli przekonania nie będą dosłownie przechowywane w innym miejscu niż pragnienia, to i tak będą musiały (1) być tożsame z jakimiś wewnętrznymi obliczeniowymi stanami czy komponentami systemu poznawczego oraz (2) w jakiś sposób różnić się strukturalnie od pragnień (w przeciwnym razie nie będą mogły w ogóle być „odróżniane" od pragnień w systemie; por. Piccinini, Craver 2011).

konania o własnej śmierci, rozkładzie własnego ciała albo o znajdowaniu się w piekle); (2) dysfunkcję systemu ewaluacji przekonań, który normalnie pozwoliłby na odrzucenie urojeniowego przekonania na podstawie dostępnych świadectw (Coltheart 2005; Coltheart, Langdon, McKay 2011). Zwróćmy uwagę na system ewaluacji przekonań wymieniony w punkcie (2). Ewaluacja przekonań to proces występujący na poziomie osobowym. Ma on charakter racjonalny i polega na nabywaniu, modyfikowaniu oraz odrzucaniu przekonań w sposób maksymalizujący ich spójność i poziom uzasadnienia. Kiedy przyjrzymy się jednak literaturze dotyczącej urojeń, okaże się, że postulowany system ewaluacji przekonań stanowi strukturę o charakterze subosobowym. Jest on zlokalizowany w ośrodkowym układzie nerwowym i zajmuje się racjonalną rewizją przekonań na temat świata, a zatem tym, co wydaje się stanowić domenę podmiotów intencjonalnych. Neuronalną lokalizację tego systemu można odkryć za pomocą badań wykorzystujących neuroobrazowanie oraz analizę poznawczych skutków lezji w określonych obszarach korowych (Coltheart, Langdon, McKay 2011). Innymi słowy, zgodnie z taką teorią gdzieś w ludzkim mózgu znajduje się system przechowujący czy kodujący stany przekonaniowe oraz racjonalnie rewidujący je w odniesieniu do stosownych świadectw i wzajemnej koherencji. Propozycja ta stanowi więc dobrą ilustrację praktycznego zastosowania ZKOS w kognitywistyce. Tak jak w przypadku diagramu Sticha i Nicholsa, tak i tu uwidacznia się postrzeganie systemu poznawczego na poziomie subosobowym jako „skrzynki" na postawy propozycjonalne. W dalszej części tego rozdziału chcę zaproponować alternatywę dla tej wizji.

5.1.3. Jak rozumieć relację między poziomem osobowym a subosobowym? O potrzebie alternatywy dla założenia korespondencji

Zapytajmy najpierw, dlaczego mielibyśmy w ogóle szukać alternatywnego wobec ZKOS sposobu rozumienia relacji między poziomem osobowym a subosobowym? Co stanowi teoretyczną motywację do takich poszukiwań? Należy wyraźnie zaznaczyć, że moim

celem nie jest wskazanie argumentu jednoznacznie dyskwalifikującego założenie korespondencji i przez to niejako „zmuszającego" do odnalezienia alternatywy. Przyznaję, że nie mam takiego argumentu. Zamiast tego chcę (1) wskazać, dlaczego ZKOS może wydawać się dla współczesnych filozofów stanowiskiem uzasadnionym, atrakcyjnym, a nawet jedynym możliwym do racjonalnego utrzymywania (w każdym razie dla naturalistów), a następnie (2) wykazać, że wbrew pozorom ZKOS nie jest pozycją tak uzasadnioną i atrakcyjną, jak się na ogół wydaje, a na pewno nie jest jedyną możliwą do racjonalnego utrzymywania koncepcją relacji między poziomem osobowym a subosobowym. Właśnie dlatego poszukiwanie alternatywnego, doskonalszego – lepiej uzasadnionego, bardziej realistycznego empirycznie – ujęcia relacji międzypoziomowych jest jak najbardziej pożądane.

Skąd zatem założenie korespondencji czerpie swój status niemal niekwestionowanej presupozycji? Sądzę, że istnieją cztery tego rodzaju źródła. Pierwsze z nich wiąże się z postrzeganiem ZKOS jako założenia „wbudowanego" w pojęciowe zobowiązania psychologii potocznej. Drugie – z postrzeganiem ZKOS jako dobrze ugruntowanej, wiarygodnej hipotezy empirycznej. Trzecie – z pewnymi presupozycjami mającymi rodowód w filozofii nauki, a dokładniej w nomologiczno-dedukcyjnym modelu wyjaśniania. Czwarte źródło jest związane z postrzeganiem założenia korespondencji jako jedynej naturalistycznej alternatywy dla antynaturalistycznych stanowisk głoszących całkowitą autonomię poziomu osobowego względem subosobowego. Jak postaram się wykazać, każda z tych podstaw do utrzymywania ZKOS jest chwiejna.

a) ZKOS jako wyraz pojęciowych zobowiązań architekturalnych psychologii potocznej

Ktoś może stwierdzić, że założenie korespondencji stanowi bezpośrednią konsekwencję istnienia *pojęciowych* (pojęciowo czy analitycznie ugruntowanych) związków między aparaturą konceptualną psychologii potocznej a określonymi twierdzeniami dotyczącymi subosobowej architektury poznania. Z takiej perspektywy potoczne pojęcia osobowych stanów intencjonalnych miałyby konceptual-

nie „wbudowane" założenia czy twierdzenia dotyczące wewnętrznej budowy systemu poznawczego. Na przykład stanowiłoby konieczną prawdę pojęciową – wynikającą z treści (intensji) potocznego pojęcia przekonania – iż bycie przekonanym, że *p*, polega na posiadaniu „w głowie" pewnej subosobowej struktury o określonych własnościach intencjonalnych (kodującej treść, że *p*) oraz funkcjonalnych (pełniącej role przyczynowe odpowiadające rolom przypisywanym przekonaniu, że *p*).

Przyjrzyjmy się z takiej perspektywy eliminatywizmowi architekturalnemu. Aby utrzymywać swoją pozycję, jego zwolennik jest zmuszony twierdzić, że ZKOS wynika z pojęciowo ugruntowanych zobowiązań architekturalnych psychologii potocznej (Horgan, Graham 1991; Horgan 1993; Henderson, Horgan 2004). Jedynie pod warunkiem przyjęcia takiego założenia eliminatywiści architekturalni mogą twierdzić, że nieistnienie subosobowych odpowiedników postaw propozycjonalnych stanowi dobrą rację za odrzuceniem istnienia postaw. Gdyby było inaczej – na przykład gdyby ZKOS stanowiło pewną hipotezę *empiryczną*, a nie było ugruntowane pojęciowo – to argument za eliminatywizmem architekturalnym (sformułowany wedle schematu przedstawionego w sekcji 5.1.2) byłby zupełnie niekonkluzywny. Przejście od przesłanek do wniosku mogłoby zostać łatwo zablokowane twierdzeniem, że istnienie postaw propozycjonalnych nie wymaga istnienia subosobowych odpowiedników osobowych stanów intencjonalnych, w związku z czym nieistnienie takich odpowiedników nie stanowi żadnej racji za nieistnieniem postaw (por.: Horgan, Graham 1991; Horgan 1993; Henderson, Horgan 2004)[13]. Wykonanie takiego uderzającego w eliminatywizm architekturalny ruchu jest jednak niemożliwe, o ile założymy, iż twierdzenia architekturalne są nieodzowną i konieczną, *konceptualnie* „wbudowaną" częścią potocznych pojęć mentalnych. Jeśli przyjmiemy takie założenie, to zachodzenie ści-

[13] Zamiast odrzucać twierdzenie o istnieniu przekonań i pragnień, moglibyśmy w takiej sytuacji zrewidować hipotezę empiryczną, zgodnie z którą przekonania i pragnienia są identyczne z pewnymi strukturami subosobowymi. Gdybyśmy zdecydowali się na taki krok, z faktu nieistnienia subosobowych odpowiedników przekonań i pragnień nie wynikałoby nieistnienie przekonań i pragnień.

śle określonych faktów architekturalnych będzie (*a priori*) konieczne dla istnienia przekonań i pragnień. Oponent eliminatywisty architekturalnego nie będzie miał pola manewru. Podsumowując, argument za tym stanowiskiem wydaje się „działać", tylko jeśli potraktujemy ZKOS jako wyraz zobowiązań teoretycznych „wbudowanych" w potoczne pojęcia mentalne.

Istnieje jednak kilka racji za odrzuceniem idei, jakoby wywodzące się z psychologii potocznej pojęcia postaw propozycjonalnych niosły ze sobą zobowiązania dotyczące subosobowej architektury poznania. David Henderson i Terrence Horgan (2004) formułują szereg eksperymentów myślowych, które odwołują się do intuicji semantycznych związanych z pojęciami postaw propozycjonalnych. Eksperymenty te pokazują, że nasze potoczne intuicje semantyczne[14] świadczą przeciwko tezie o konceptualnym związku między psychologią potoczną a twierdzeniami dotyczącymi subosobowej organizacji systemu poznawczego[15]:

– **Eksperyment 1** (argument z upartych intuicji). Wyobraźmy sobie, że ktoś wykazuje nam ponad wszelką wątpliwość, iż subosobowa architektura systemów poznawczych nie zawiera struktur będących funkcjonalnymi/intencjonalnymi odpowiednikami postaw propozycjonalnych. Czy w takiej sytuacji byłoby dla nas naturalne uznać, że nikt nigdy niczego nie pragnął i nie był

[14] Intuicje semantyczne mają tu grać rolę „okna" dającego nam wgląd w treść naszych potocznych pojęć mentalnych (por. Goldman 2007). Trzeba podkreślić, że argumentacja Hendersona i Horgana nie wymaga uznania, iż intuicje semantyczne świadczą o czymkolwiek więcej niż tylko o treściach pojęć (rozumianych psychologicznie), którymi się posługujemy. Eksperymenty te mają nam dać wgląd w to, jak *konceptualizujemy* stany mentalne, a nie w *naturę* tych stanów. (Zakładam tu jednocześnie treściową homogeniczność pojęć mentalnych we wspólnotach posługujących się psychologią potoczną. To znaczy , że eksperymenty myślowe Hendersona i Horgana generowałyby zbliżone intuicje wśród osób wykorzystujących pojęcia osobowych stanów intencjonalnych, niezależnie od ich bagażu kulturowego czy socjoekonomicznego. Jest to rzecz jasna hipoteza, która docelowo wymaga potwierdzenia, na przykład za pomocą stosownych badań z zakresu filozofii eksperymentalnej).

[15] Eksperymenty Hendersona i Horgana są tu opisane w sposób zaadaptowany do aparatury terminologicznej stosowanej w tej pracy.

o niczym przekonany? Czy w takim scenariuszu przypisywanie ludziom postaw propozycjonalnych wydałoby nam się nagle semantyczną czy pojęciową pomyłką? Wygląda na to, że nie. Gdyby jednak ZKOS miało być wbudowane *pojęciowo* w psychologię potoczną, powinniśmy oczekiwać, że odpowiedzi na te pytania będą twierdzące.

– Eksperyment 2 (argument z konserwatyzmu pojęciowego). Wiele pojęć o zasadniczym znaczeniu dla ludzkiego życia społecznego (stwierdzanie, posiadanie racji, uzasadnianie, działanie, odpowiedzialność) presuponuje adekwatność opisu ludzkich działań w kategoriach psychologii potocznej. Wiele ważnych aspektów naszego życia społecznego opiera się na założeniu, że ludzie działają na podstawie przekonań i pragnień. Gdyby więc psychologia potoczna była rzeczywiście powiązana pojęciowo z ZKOS, mogłoby to – w sytuacji, gdyby okazało się, że międzypoziomowa korespondencja nie zachodzi – iść pod prąd istotnym zastosowaniom psychologii potocznej. Aby psychologia potoczna mogła spełniać swoje zadania, powinna być wolna od tego rodzaju nadmiernie restrykcyjnych zobowiązań pojęciowych. Jednocześnie kryteria tego, jak używamy pojęć, mają charakter w dużym stopniu pragmatyczny. Oznacza to między innymi, że unikamy posługiwania się pojęciami w sposób na tyle restrykcyjny, iż mogłoby to zagrażać funkcji, jakie te pojęcia pełnią w naszym życiu. Tym samym jest mało prawdopodobne, by psychologia potoczna była powiązana pojęciowo z tezami, które mogą w zasadniczy sposób zagrozić jej użyteczności w ludzkim życiu społecznym. Gdybyśmy stanęli przed wyborem między zachowaniem a odrzuceniem osobowych pojęć mentalnych pod wpływem rozwoju kognitywistyki, należałoby wybrać – ze względów pragmatycznych – tę pierwszą opcję.

– Eksperyment 3 (argument z asymetrycznej wyobrażalności). Nie mamy problemów z wyobrażaniem sobie, że odrzucamy jako fałszywą jakąś kognitywistyczną teorię postulującą istnienie korespondencji między poziomem osobowym a subosobowym

(na przykład Fodorowską koncepcję języka myśli). Jednak wyobrażenie sobie, że odrzucamy kategorie psychologii potocznej – dajmy na to pojęcie przekonania czy celowego działania – jest bardzo trudne lub wręcz niemożliwe. Wydaje się, że porzucenie pojęcia celowego działania samo musiałoby być rozumiane jako celowe działanie, a porzucenie pojęcia przekonania samo musiałoby być konceptualizowane jako „przekonanie o nieistnieniu przekonań". Taka asymetria w wyobrażalności obydwu scenariuszy – pierwszego, w którym odrzucamy określoną teorię naukową, oraz drugiego, w którym odrzucamy elementy psychologii potocznej – naturalnie sprzyja twierdzeniu, że nie istnieje konieczny związek pojęciowy między psychologią potoczną a faktami dotyczącymi subosobowej architektury. Gdyby bowiem psychologia potoczna niosła ze sobą pojęciowe zobowiązania architekturalne, obie możliwości powinny być równie łatwo (lub równie trudno) wyobrażalne.

Do eksperymentów myślowych Hendersona i Horgana warto dodać sformułowany przez Frances Egan (1995) argument odwołujący się do świadectw empirycznych. Autorka ta powołuje się na ustalenia psychologii rozwojowej i zwraca uwagę na fakt, że już pięcioletnie dzieci dysponują pojęciami przekonania, pragnienia czy intencji[16]. Nie istnieją jednak żadne powody, by sądzić, że tak małe dzieci mają jakiekolwiek pojęcie o tym, jak przekonania, pragnienia czy intencje są (lub mogą być) realizowane czy implementowane na poziomie subosobowym. Co więcej, pojęcia stanów mentalnych posiadane przez osoby dorosłe nie są zasadniczo odmienne od pojęć nabytych przez nich w dzieciństwie. Na przykład „dorosłe" pojęcie intencji jest oparte na „dziecięcym". Skoro nabywane w dzieciństwie pojęcia mentalne nie mają konceptualnie „wbudowanych" zobowiązań architekturalnych, możemy przyjąć, że podobnych zobowiązań nie mają także najprawdopodobniej pojęcia „dorosłe".

[16] Warto dodać, że niektóre nowsze, choć czasem dość kontrowersyjne, badania sugerują, iż dzieci wykazują znajomość niektórych pojęć mentalnych jeszcze wcześniej, nawet w wieku kilku lub kilkunastu miesięcy (por.: Gergely, Csibra 2003; Onishi, Baillargeon 2005; Csibra 2008; He, Bolz, Baillargeon 2011).

Nie wydaje się zatem, by potoczne pojęcia postaw propozycjonalnych były konceptualnie powiązane z jakimikolwiek tezami dotyczącymi wewnętrznej, subosobowej architektury systemów poznawczych. Wszystkie przytoczone argumenty stanowią mocną rację za odrzuceniem twierdzenia, iż ZKOS jest ugruntowane w pojęciowych zobowiązaniach psychologii potocznej.

b) ZKOS jako hipoteza empiryczna

Jaki inny status może mieć założenie korespondencji, jeśli nie jest ono ugruntowane w pojęciowych zobowiązaniach psychologii potocznej? Jak już zostało wspomniane, wydaje się, że można potraktować ZKOS jako hipotezę empiryczną. Przy takiej interpretacji ZKOS jest empirycznie falsyfikowalną hipotezą głoszącą, że istoty posiadające przekonania i pragnienia to *de facto* systemy poznawcze spełniające określone subosobowe warunki architekturalne. Te ostatnie są wyrażone właśnie w (metafizycznie sformułowanym) założeniu korespondencji.

Podejście tego rodzaju *explicite* akceptuje chociażby Martin Davies (1998; 2000). Wychodzi on od inspirowanej pracami Christophera Peacocke'a charakterystyki osobowych procesów inferencyjnych jako ucieleśniających pewne ukryte (*tacit*) reguły formalne[17]. Następnie Davies przyjmuje, że każdy proces, który rzeczywiście jest oparty na ukrytej formalnej regule, musi cechować się „systematycznością przyczynową". Oznacza to, że każda pojedyncza aplikacja danej reguły powinna mieć to samo źródło przyczynowe (w przeciwnym razie nie mielibyśmy do czynienia z przestrzeganiem reguły, a działaniem w sposób, który może być opisany, tak *jak gdyby* ktoś przestrzegał tej reguły). Także ukryte reguły myślenia na poziomie osobowym powinny mieć „systematyczne" źródła przyczynowe. Źródła te są zlokalizowane na poziomie subosobowym. Na przykład za każdym razem, gdy ktoś aplikuje określoną regułę inferencyjną, za aplikację tę odpowiada jeden i ten sam subosobowy stan obli-

[17] Jest to dość słaby punkt w całej argumentacji Daviesa, ponieważ wymaga przyjęcia bardzo mocnych filozoficznych założeń dotyczących poziomu osobowego (Skidelsky 2006). Nie będę tu jednak polemizować z tym założeniem.

czeniowy. Ze względu na formalny charakter osobowych reguł inferencyjnych Davies przyjmuje, że u ich podstaw musi stać syntaktycznie ustrukturyzowany, *quasi*-lingwistyczny kod, słowem – język myśli. Procesy obliczeniowe realizowane w takim subosobowym kodzie podzielają szereg ważnych własności z osobowymi procesami inferencyjnymi. Tym samym dla tego autora opis zjawisk mentalnych na poziomie osobowym stanowi punkt wyjścia do postawienia hipotezy o subosobowych podstawach poznania. Zgodnie z tą hipotezą osobowe procesy inferencyjne są odzwierciedlone w subosobowych procesach obliczeniowych.

Zauważmy przede wszystkim, że powyższe rozumowanie nie jest argumentem transcendentalnym. Davies nie określa „warunków możliwości" bycia podmiotem przekonań i pragnień. Wydaje się raczej, że autor ten dochodzi do ZKOS na podstawie rozumowania abdukcyjnego, stwierdzając, że teza o międzypoziomowej korespondencji stanowi najlepsze wyjaśnienie określonych fenomenów z poziomu osobowego, w szczególności zaś zdolności do wykonywania racjonalnych procesów inferencyjnych[18].

Zwróćmy uwagę na dwie istotne konsekwencje potraktowania ZKOS jako hipotezy empirycznej. Po pierwsze, przy takim postawieniu sprawy okazuje się, że możemy odrzucić założenie korespondencji, nie odrzucając jednocześnie tezy o tym, że świat zamieszkują istoty działające na podstawie posiadanych postaw propozycjonalnych. Zamiast rewidować przekonanie o istnieniu przekonań i pragnień, możemy zrewidować przekonanie o tym, jaka relacja *de facto* zachodzi między poziomem osobowym a subosobowym. ZKOS okazuje się tylko jedną z potencjalnie wielu konkurujących hipotez naukowych. Otwiera się przed nami możliwość, że istnieją koncepcje relacji między poziomem osobowym a subosobowym, które (1) nie

[18] Prezentuję tu tylko część stanowiska Daviesa w sprawie relacji między poziomem osobowym a subosobowym. Warto wspomnieć, że oprócz tego rodzaju rozumowań „w dół" – gdzie przesłanki dotyczą poziomu osobowego, a wnioski subosobowego – Davies (2000) postuluje także istnienie luk eksplanacyjnych idących „w górę", z poziomu subosobowego do osobowego. Za przykład takiej luki uznaje on niemożność wyjaśnienia świadomości fenomenalnej w kategoriach subosobowych.

postulują zachodzenia korespondencji międzypoziomowej, (2) nie niosą eliminatywistycznych konsekwencji (jeśli takowa korespondencja nie zachodzi), (3) są lepiej niż ZKOS uzasadnione w świetle świadectw empirycznych i wiedzy dotyczącej poziomu subosobowego. Sam wspomniany już Davies (1998) nie wyklucza możliwości, że hipoteza języka myśli okaże się fałszywa. Jak sam przyznaje, gdyby teoria ta rzeczywiście okazała się fałszywa, to nie tyle zagrażałoby to realizmowi względem zjawisk z poziomu osobowego, ile raczej kazałoby zrewidować nasz pogląd, jeśli chodzi o subosobowe podstawy tych zjawisk. Musielibyśmy wtedy dokonać „pojęciowych negocjacji" w sprawie architekturalnych zobowiązań psychologii potocznej.

Po drugie, założenie korespondencji – potraktowane nawet jako hipoteza empiryczna – nie wydaje się stać na mocnych podstawach. Pomijam tu już głosy przekonujące, że kognitywistyka tak naprawdę nigdy, nawet w czasach dominacji GOFAI nie była do pogodzenia z ZKOS (por.: Stich 1983; Matthews 2007: 36–96; Ramsey 2007). Skupmy się tylko na nowszych trendach teoretycznych w naukach kognitywnych. Otóż wydaje się, że kognitywistyka w trakcie ostatnich dekad coraz bardziej oddala się od postrzegania architektury poznania w sposób, który pozwalałby na utożsamianie wewnętrznych komponentów (stanów komponentów) systemu poznawczego z postawami propozycjonalnymi (por. m.in.: Ramsey, Stich, Garon 1990; Beer 2003; Świątczak 2003; Ramsey 2007: 203–235). Co obecnie jak najbardziej prawdopodobne, wizja subosobowej architektury systemu poznawczego zawarta w dojrzałej kognitywistyce nie będzie postulować stanów czy struktur posiadających funkcjonalne i intencjonalne własności postaw propozycjonalnych. Bardzo możliwe na przykład, że w kognitywistyce będzie rosnąć znaczenie podejścia dynamicznego (Ramsey 2007: 203–222). Jeśli tak się stanie, zapewne odejdziemy od wizji „skrzynki na postawy propozycjonalne" jeszcze bardziej, niż stało się to po powstaniu koneksjonizmu. Być może zamiast trwać w nadziei na utrzymanie ZKOS, powinniśmy jako filozofowie poszukać alternatyw, koncepcji zrywających z ideą suboso-

bowych odpowiedników stanów osobowych[19]. Scenariusz, w którym musimy dokonać „pojęciowych negocjacji" dotyczących architekturalnych zobowiązań psychologii potocznej, o którym Davies jedynie spekulował, wydaje się ziszczać na naszych oczach.

c) ZKOS jako konsekwencja założeń dotyczących natury wyjaśniania i redukcji w nauce

Problemy z ZKOS jako hipotezą empiryczną nie kończą się jednak na prawdopodobnych, choć w pewnym stopniu jedynie spekulatywnych scenariuszach dotyczących rozwoju nauk kognitywnych.

[19] Nie twierdzę, że nikt takich alternatyw jeszcze nie sformułował. Chodzi raczej o to, że tego rodzaju alternatywne teorie nie wyznaczają w filozofii umysłu głównego nurtu i są nadal postrzegane jako niosące zagrożenie „instrumentalizmem" czy „behawioryzmem" względem postaw propozycjonalnych. Jak postaram się dalej pokazać, prezentowana tu teoria poziomu osobowego oraz postaw propozycjonalnych zrywa z założeniem korespondencji, lecz nie jest przy tym ani instrumentalistyczna, ani behawiorystyczna (w każdym razie nie w sensie behawioryzmu analitycznego). Dobrym przykładem innej, obecnej w literaturze teorii, która również zrywa z założeniem korespondencji – a której czasem zarzuca się (sądzę, że niesłusznie) właśnie instrumentalizm i behawioryzm – jest koncepcja Daniela Dennetta (2003, 2008). Prace Dennetta znacząco wpłynęły zresztą na moje myślenie o relacji między psychologią potoczną a kognitywistyką. Omówienie analogii i różnic między bronionym tu przeze mnie stanowiskiem a ideami tego autora to bardzo szeroki temat, którego pełne omówienie zbyt mocno odwiodłoby mnie od głównego wątku rozważań. Warto jednak wskazać dwa fundamentalne, jak sądzę, punkty zbieżne. Po pierwsze, przyjmowane tu przeze mnie twierdzenia o różnicy funkcji eksplanacyjnych pełnionych przez psychologię potoczną i kognitywistykę są zbliżone do rozróżnienia Dennetta na wyjaśnienia stworzone za pomocą: strategii „intencjonalnej" (odpowiednik psychologii potocznej) oraz strategii „projektowej" czy „funkcjonalnej" (odpowiednik mechanistycznych wyjaśnień kognitywistyki). O ile jednak Dennett (2003) podkreśla znaczenie, jakie dla strategii intencjonalnej ma założenie racjonalności wyjaśnianego systemu, o tyle broniona tu koncepcja traktuje wyjaśnienia intencjonalne jako przyczynowe, lecz skoncentrowane na systemowym i „ekologicznym" poziomie organizacji (por. jednak Gładziejewski 2012). Po drugie, podobnie jak Dennett, będę tu przyjmował perspektywę, zgodnie z którą bycie podmiotem postaw propozycjonalnych to kwestia wykazywania wzorców działania, które są „widoczne" tylko z perspektywy psychologii potocznej (czyli strategii intencjonalnej). W tej pracy będę się jednak starał dookreślić tę ideę w sposób odbiegający od propozycji Dennetta, wykorzystując dyspozycyjną teorię postaw propozycjonalnych.

W rozdziale 2 (sekcja 2.1.2) zaznaczałem, że sposób myślenia o problemie (problemach) reprezentacji mentalnych w analitycznej filozofii umysłu został w jakimś zakresie ukształtowany przez założenia pochodzące z filozofii nauki. Warto teraz rozwinąć tę myśl.

Wiele znaczących filozoficznych sporów dotyczących reprezentacji koncentrowało się wokół zagadnienia, czy możliwa jest *redukcja interteoretyczna* psychologii potocznej do kognitywistyki (Fodor 1975, 1987, 2001; Churchland 1981; Stich 1983). Tego rodzaju podejście wyrażają chociażby prace Fodora (1975, 1987, 2001). Autor ten przyjmuje, że: (1) psychologia potoczna stanowi teorię specyfikującą *prawa* rządzące (osobowymi) stanami intencjonalnymi; (2) aby zachować realizm względem bytów postulowanych przez psychologię potoczną (to znaczy potocznych, osobowych stanów intencjonalnych) oraz względem formułowanych za jej pomocą wyjaśnień, powinniśmy ją naukowo „zrehabilitować". Na czym polegałaby jednak w takiej perspektywie realizacja punktu (2)? To rzecz jasna szeroki wątek z zakresu egzegezy Fodora, jednak skupmy się tu na jednym z wątków, które należałoby podjąć, odpowiadając na to pytanie. Otóż wedle wymienionego autora powinniśmy być zdolni do pokazania, że intencjonalne prawa psychologii potocznej są poprawne w świetle prawdziwej naukowej teorii umysłu. Chodzi jednak nie o teorię neuronaukową – tu powstawałby problem związany z wieloraką realizowalnością rodzajów intencjonalnych przez rodzaje neurobiologiczne – lecz z zakresu *psychologii obliczeniowej*. Aby zatem naukowo zrehabilitować psychologię potoczną, powinniśmy pokazać, że prawa przez nią postulowane dają się *zredukować* do tych rządzących czysto „syntaktycznymi" elementami obliczeniowej architektury umysłu[20]. Jeśli postulaty Fodora są poprawne, powinniśmy być w stanie wydedukować prawa tej pierwszej teorii z praw dru-

[20] Sam Fodor, mówiąc o relacjach interteoretycznych, często posługuje się neutralnym terminem „implementacja", stwierdzając także, iż prawa teorii z wyższego poziomu są implementowane przez prawa wyższego poziomu (por. np. Fodor 2001). Jak sam jednak zauważa, implementacja zawsze oznacza albo redukcję, albo wieloraką realizację (i przez to nieredukowalność). Z punktu widzenia Fodora (1975; 1987; 2001) relacja między prawami intencjonalnymi a neuronaukowymi polega na wielorakiej realizacji. Jednocześnie wkłada on sporo wysiłku

giej. Takie podejście rodzi rzecz jasna pytanie: co należy począć, jeśli tego rodzaju redukcja interteoretyczna nie powiedzie się? Otóż inni filozofowie – przejmując (bardziej lub mniej otwarcie oraz często z pewnymi modyfikacjami) wymienione założenia Fodora – bronili tezy, że „pokojowa" redukcja interteoretyczna psychologii potocznej do kognitywistyki jest niemożliwa. Na tej podstawie formułowali oni twierdzenie, że kategorie psychologii potocznej będą musiały zostać wyeliminowane ze słownika nauki (por.: Churchland 1981; Stich 1983; Ramsey 2007; por. także przypis 8).

Sposób rozumienia teorii, relacji interteoretycznych, wyjaśniania oraz redukcji w wymienionych koncepcjach jest bardzo zbliżony do tego (lub wręcz identycznym z tym), co o wszystkich tych zagadnieniach mówi model nomologiczno-dedukcyjny (N-D). Szczególnie klarownie widać to w teorii Fodora. Psychologia potoczna okazuje się teorią złożoną z praw. Jest ona teorią redukowaną. Natomiast kognitywistyka – a w każdym razie jej część zmierzająca do odkrycia *obliczeniowych* podstaw umysłu – ma dostarczyć teorię redukującą. Procedura redukcji polega na dedukcyjnym wywiedzeniu praw teorii redukowanej z praw teorii redukującej, *via* prawa mostowe łączące słowniki obu teorii (czyli łączące intencjonalny słownik psychologii potocznej z czysto syntaktycznym słownikiem psychologii obliczeniowej czy kognitywistyki)[21]. Projekt naturalizacji intencjonalności może być z takiej perspektywy zinterpretowany jako zmierzający do pokazania, na czym polega, czy też dzięki czemu jest możliwa prawdziwość praw mostowych łączących kategorie intencjonalne psychologii potocznej z kategoriami kognitywistyki. W przypadku wspomnianego Fodora ten projekt miałby dostarczyć

w pokazanie, że między prawami intencjonalnymi a prawami psychologii *obliczeniowej* może zachodzić relacja *redukcji*.

[21] Oczywiście powstaje tu problem dotyczący tego, jak dokładnie należy interpretować prawa mostowe (por. Poczobut 2009: 146–159). Wydaje się, że powinny one być rozumiane zdecydowanie mocniej, niż jako jedynie stwierdzające istnienie korelacji między przekonaniami i pragnieniami a stanami subosobowymi. Na przykład projekt Fodora wymaga zachodzenia identyczności typów między bytami postulowanymi przez psychologię potoczną a bytami postulowanymi przez kognitywistykę (psychologię obliczeniową).

odpowiedzi na pytanie: jak czysto obliczeniowe struktury podlegające prawom kognitywistyki (psychologii naukowej/obliczeniowej) mogą być (rodzajowo) identyczne z intencjonalnymi stanami postulowanymi przez psychologię potoczną?[22]

Powyższe obserwacje wiążą się z założeniem korespondencji w następujący sposób. Opisany wyżej sposób postrzegania relacji interteoretycznych między psychologią potoczną a kognitywistyką naturalnie sprzyja akceptacji ZKOS. Zwróćmy uwagę na rolę praw mostowych w opisywanym tu rodzaju redukcji. Tworzą one konceptualny most między teorią redukowaną a redukującą, w naszym przypadku – między psychologią potoczną a kognitywistyką. Jeśli przyjmiemy, że ontologia pierwszej teorii obejmuje stany intencjonalne (i procesy inferencyjne, którym one podlegają) o charakterze osobowym, a ontologia tej drugiej obejmuje stany (fakty, struktury, procesy) subosobowe, to powiązanie słowników obu teorii będzie skutkowało przyjęciem właśnie ZKOS (lub stanowiska bardzo zbliżonego). Prawa mostowe będą wszakże pojęciowo przyporządkowywać byty, których dotyczą prawa psychologii potocznej – bytom, których dotyczą prawa teorii kognitywistycznej. Skuteczna redukcja wymaga z takiej perspektywy zachodzenia międzypoziomowej korespondencji. Istnieje zatem związek między ZKOS a założeniami z zakresu filozofii nauki (por. też omówienie wpływu modelu N-D na filozofię umysłu w: Bechtel 2008). Nie twierdzę, że *wszyscy* filozofowie przyjmujący ZKOS podzielają te założenia, jednak stawiam hipotezę, że założenie korespondencji przynajmniej częściowo zawdzięcza swoją powszechność powszechności określonych założeń pochodzących z filozofii nauki.

Sądzę, że związek ZKOS z modelem N-D świadczy w istocie przeciwko temu pierwszemu. Dzieje się tak z dwóch powodów. Po pierwsze, jak już wcześniej zostało wspomniane (sekcja 3.3.2), model

[22] Pytanie to stało się dla Fodora szczególnie kłopotliwe po tym, jak zdecydował się on na akceptację semantyki informacyjnej (por. Fodor 2001). Obliczeniowe operacje są czułe jedynie na formalne, a nie informacyjne własności struktur, które w tych operacjach uczestniczą. Jak ujmuje to sam Fodor: „Wygląda na to, [...] że nie mogę utrzymać jednocześnie semantyki informacyjnej i psychologii obliczeniowej. Jest to niezwykle irytujące, bo bardzo zależy mi na obu" (2001: 26).

N-D wyjaśniania oraz związana z nim wizja redukcji są krytykowane jako w znacznym stopniu deskryptywnie nieadekwatne w odniesieniu do kognitywistyki (Godfrey-Smith 2005; Craver 2007: 228–271; Bechtel 2008: 142–157). Jeśli uznamy zatem wywodzącą się z modelu N-D koncepcję redukcji interteoretycznej za ważną teoretyczną motywację do przyjęcia ZKOS, to wskazując nieadekwatność takiej koncepcji redukcji, osłabiamy zarazem ZKOS. Mówiąc inaczej, okazuje się, że skuteczne wyjaśnianie redukcyjne w kognitywistyce nie wymaga, aby założenie korespodencji (w metafizycznym sformułowaniu) było prawdą. Po drugie, z perspektywy bardziej realistycznego modelu wyjaśniania w kognitywistyce – to znaczy modelu mechanistycznego – nie ma dobrych podstaw, aby oczekiwać, iż zachodzi korespondencja między poziomem osobowym a subosobowym. Wyjaśnianie w kognitywistyce to przede wszystkim odkrywanie i opisywanie mechanizmów, a nie dedukcja z praw. Jednak z mechanistycznej perspektywy, własności wyższego rzędu, czyli własności systemowe, są egzemplifikowane przez *systemy jako całości*, nie przez byty z *niższych poziomów organizacji*. Jak zobaczymy w podrozdziałach 5.2. i 5.3, przyjęcie takiej mechanistycznej optyki sprzyja temu, by ująć relację zachodzącą między poziomem osobowym a subosobowym w sposób całkowicie zrywający z ZKOS.

d) ZKOS jako jedyna opcja dla naturalisty

W sekcji 5.1.1 wspomniałem, że osobowy poziom wyjaśniania jest do pewnego stopnia autonomiczny względem subosobowego. Mam przez to na myśli, że jest możliwe formułowanie kompletnych i poprawnych wyjaśnień osobowych, które nie odwołują się w ogóle do faktów o charakterze subosobowym. Można argumentować, że taka autonomia kończy się dopiero wtedy, gdy dochodzi do wyjaśniania działań radykalnie nieracjonalnych czy chaotycznych, na przykład towarzyszących niektórym chorobom psychicznym. Dostarczenie kompletnych wyjaśnień takich działań na poziomie osobowym bywa niemożliwe, w związku z czym może występować konieczność odwołania się do poziomu subosobowego (Bermúdez 2000, 2005: 17–39).

Jak jednak zauważa Jose Luiz Bermúdez (2000, 2005: 40–52), istnieje też dużo mocniejszy sposób pojmowania autonomii poziomu osobowego. Dla zwolenników takiego podejścia wyjaśnienia na obu poziomach są radykalnie niewspółmierne (por.: Davidson 1992[23]; Hornsby 2000). Z takiego punktu widzenia oba poziomy są tak różne i niezależne, że nie istnieje nic, co pozwalałoby na wystąpienie jakichkolwiek pojęciowych czy eksplanacyjnych związków między nimi. Nie mogą tu więc zachodzić żadne teoretycznie ciekawe interakcje. Wyjaśnienia subosobowe są „na inny temat" niż te osobowe. Jako źródło owej niewspółmierności wskazuje się czasem fakt, że wyjaśnienia na poziomie osobowym posiadają cechę całkowicie nieobecną na poziomie subosobowym: normatywność. Postawy propozycjonalne są nie tylko przyczynami działań, ale też ich *racjami*. Wyjaśniają one działania nie (tylko) jako ich przyczyny, ale (też) jako ich racje. O ile wyjaśnienia subosobowe mają czysto deskryptywny charakter, o tyle osobowe w sposób nieunikniony odwołują się do normatywnych ideałów racjonalności i spójności (Hornsby 2000; por. Bermúdez 2000, 2005: 40–52).

Przyjęcie tak mocnej autonomii poziomu osobowego prowadzi w istocie do zerwania z naturalizmem jako takim. Negując istnienie jakichkolwiek ciekawych związków konceptualnych i eksplanacyjnych między poziomem osobowym a subosobowym, całkowicie rezygnuje się jednocześnie z Sellarsowskiego projektu uzgodnienia dwóch obrazów świata. Zamiast szukać synoptycznej wizji uzgadniającej obraz naukowy z manifestującym się, przystajemy wtedy na dualizm. Nie jest to dualizm metafizyczny, lecz epistemiczny, postulujący istnienie dwóch radykalnie różnych i niewspółmiernych sposobów opisywania i wyjaśniania umysłu czy poznania. Takie ujęcie gwarantuje badaniom i analizom prowadzonym na poziomie osobowym – w tym czysto apriorycznym analizom filozoficznym – całkowitą niezależność względem kognitywistycznych dociekań dotyczących subosobowej organizacji systemu poznawczego.

[23] Co prawda sam Davidson nie posługuje się dystynkcją osobowe–subosobowe, jednak jego monizm anomalny implikuje tezę o niewspółmierności poziomu osobowego względem subosobowego (por. Bermúdez 2005: 40–52).

Dla większości współczesnych filozofów umysłu taka antynaturalistyczna i „separacjonistyczna" konkluzja jest nie do przyjęcia (por. obszerną dyskusję nad tym zagadnieniem w: Miłkowski, Poczobut 2005). Co jednak stanowi alternatywę? Wydaje się, że dla wielu filozoficznych naturalistów jest nią założenie korespondencji. W kontekście ZKOS poziom osobowy okazuje się pojęciowo i eksplanacyjnie jednorodny z subosobowym. Trzeba tylko „przetłumaczyć" osobowe opisy i wyjaśnienia – na opisy i wyjaśnienia sformułowane w subosobowym języku kognitywistyki. Jedni filozofowie – ci, którzy angażują się w proces naturalizacji intencjonalności – próbują dokonać takiej translacji. Inni filozofowie – ci stojący na stanowisku eliminatywistycznym – zmierzają do pokazania, że nawet tak przetłumaczone wyjaśnienia z poziomu subosobowego okażą się fałszywe w świetle wiedzy naukowej.

Zauważmy jednak, że za przyjęcie ZKOS i pozostanie dzięki temu na stanowisku naturalistycznym trzeba zapłacić pewną filozoficzną cenę. Cenę tę stanowi całkowite pominięcie jakichkolwiek teoretycznie ważnych odmienności między poziomem osobowym a subosobowym. W sekcji 5.1.1 zostało wymienionych kilka różnic między wyjaśnieniami formułowanymi na obu tych poziomach. Wyjaśnienia osobowe dotyczą tego, *dlaczego* ludzie podejmują *poszczególne działania*; subosobowe zaś – tego, *jak* subosobowa maszyneria systemu poznawczego pozwala mu na posiadanie określonych *zdolności* poznawczych. Co więcej, wyjaśnienia osobowe są przynajmniej *częściowo* czy *względnie* autonomiczne w porównaniu do subosobowych. Przyjęcie ZKOS wydaje się całkowicie ignorować te fakty[24]. Nie tylko stany osobowe okazują się identyczne z subosobo-

[24] Drayson (2012) broni tezy, że źródłem takiego pojęciowego zamieszania jest uznanie dystynkcji osobowe–subosobowe za metafizyczną, wprowadzającą dwa rodzaje stanów mentalnych. Autorka ta twierdzi, że wspomniane rozróżnienie dotyczy (powinno dotyczyć) jedynie epistemicznych poziomów wyjaśniania. Nadanie tej opozycji interpretacji metafizycznej odbiera mu sens, bo ostatecznie prowadzi do utożsamienia stanów osobowych z jakąś subkategorią stanów subosobowych (czyli do tego, co w pracy tej określamy jako ZKOS). Zgadzam się z tezą Drayson, iż opisywana dystynkcja często nie jest odpowiednio przestrzegana w literaturze. Jednak sądzę, że wskazywana przez tę autorkę diagnoza źródeł takiego stanu rzeczy nie jest trafna. W szczególności wydaje się, że

wymi (określoną subkategorią tych stanów), ale i wyjaśnienia osobowe muszą zostać potraktowane jako „protokognitywistyczne" (por. ZKOS w sformułowaniu epistemicznym). Może to prowadzić do niepożądanych konsekwencji, na przykład: (1) traktowania kognitywistyki, tak jak gdyby jej celem było wyjaśnianie oraz przewidywanie[25] poszczególnych ludzkich *działań*, a nie wyjaśnianie zdolno-

traktowanie opozycji osobowe–subosobowe jako metafizycznej jest filozoficznie „niewinne". Po pierwsze, nie wiadomo, dlaczego mielibyśmy uznać, że próba nadania takiej metafizycznej interpretacji musi być pozbawiona sensu. Wyjaśnienia mają (na ogół) jakieś zobowiązania metafizyczne, w związku z czym jest sensowne pytanie o naturę (status metafizyczny) postulowanych przez nie eksplanansów. Prędzej czy później staniemy przed pytaniem o to, czym jest przekonanie (pragnienie, intencja i tak dalej) i jaki jest jego związek z subosobową maszynerią poznania. Problem polega nie tyle na tym, że filozofowie w ogóle interpretują dystynkcję osobowe–subosobowe jako metafizyczną, ale na tym, że interpretują ją w określony sposób, utożsamiając stany osobowe z subosobowymi. Po drugie, wielu współczesnych filozofów umysłu nie tylko dokonuje takiego utożsamienia, ale jednocześnie ignoruje różnice między strategiami eksplanacyjnymi charakterystycznymi dla obu poziomów (por. tekst główny). ZKOS jest jednym filozoficznym „pakietem", zawierającym zarówno twierdzenia dotyczące wyjaśniania, jak i twierdzenia metafizyczne, dotyczące natury postaw propozycjonalnych. Pomyłka filozoficzna związana z przyjmowaniem ZKOS – zakładając, że rzeczywiście mamy tu do czynienia z jakąś pomyłką – może wynikać nie tylko z rozstrzygnięć metafizycznych, ale także z niedocenienia różnic między osobowymi i subosobowymi wyjaśnieniami. Po trzecie, w dalszej części tego rozdziału (w podrozdziale 5.3) postaram się pokazać, że jest możliwe – wbrew twierdzeniom Drayson – nadanie opozycji osobowe–subosobowe takiej interpretacji metafizycznej, w której świetle opozycja ta może być utrzymana jako sensowna i pełniąca ważne role teoretyczne. Będę mianowicie postulować, że osobowe stany intencjonalne są własnościami wyższego rzędu systemów poznawczych (własnościami systemowymi).

[25] Można przyjąć, że zastosowanie eksplanacyjne psychologii potocznej jest ściśle powiązane z zastosowaniem predykcyjnym. Przypisywanie przekonań i pragnień pozwala nie tylko wyjaśniać działania, ale też dość precyzyjnie je przewidywać. Jak można spekulować, zdolność posługiwania się psychologią potoczną wyewoluowała – zakładając, że jest ona w ogóle skutkiem działania doboru naturalnego (por. Sterelny 2003: 211–240) – dzięki posiadanym przez siebie zastosowaniom predykcyjnym (niosącm oczywiste korzyści w dostosowaniu u wysoko uspołecznionych naczelnych), a nie eksplanacyjnym. Takie powiązanie ról predykcyjnych i eksplanacyjnych nie towarzyszy jednak (mechanistycznym) wyjaśnieniom z zakresu kognitywistyki. Dysponowanie mechanistycznym wyjaśnieniem funkcji czy zdolności pewnego złożonego systemu – w tym systemu

ści i dysfunkcji poznawczych (utożsamienie funkcji epistemicznych kognitywistyki z – epistemicznymi psychologii potocznej), czy (2) przyjmowania, że kognitywistyka wyjaśnia *zdolności* poznawcze za pomocą wewnętrznych, subosobowych odpowiedników przekonań, pragnień i innych postaw propozycjonalnych (że kognitywistyka wyjaśnia swoje eksplananda za pomocą eksplanansów charakterystycznych dla psychologii potocznej). Gdyby teza (1) była prawdziwa, to powinniśmy oczekiwać od kognitywistyki, że będzie wyjaśniać nam ona dla przykładu, dlaczego pewna osoba danego dnia sięgnęła po powieść Stanisława Lema, a nie Fiodora Dostojewskiego. Być może właśnie tak wyglądałaby w przybliżeniu kognitywistyka, gdyby koncepcje Fodora opisywały realną praktykę badawczą przedstawicieli nauk o poznaniu. Gdyby teza (2) okazała się prawdziwa, moglibyśmy oczekiwać, że kognitywistyka będzie wyjaśniać takie kompetencje, jak percepcja słuchowa czy integracja sensomotoryczna, postulując procesy inferencyjne, w których biorą udział struktury subosobowe posiadające treści intencjonalne charakterystyczne dla postaw propozycjonalnych[26]. Oba te twierdzenia wydają się bardzo problematyczne. Być może powinniśmy je jednak zaakceptować, gdybyśmy chcieli całkiem konsekwentnie ignorować dystynkcję osobowe–subosobowe.

Uważam za zasadne twierdzenie, że przywiązanie do naturalizmu stanowi dla wielu współczesnych filozofów kolejny ważny czynnik motywujący akceptację ZKOS. Dokładniej czynnikiem tym jest

poznawczego – nie wymaga i często nie przekłada się na zdolność do przewidywania zachowania tego systemu; wyjaśnienie takie pokazuje, *jak* działa ten system, lecz nie pozwala (lub pozwala w bardzo ograniczonym zakresie) przewidywać, *co* system ten *będzie* robił (por.: sekcja 2.1.2; Godfrey-Smith 2005). Jest to ciekawa rozbieżność między wyjaśnieniami osobowymi (gdzie wyjaśnianie „idzie w parze" z predykcją) a subosobowymi (gdzie predykcja i wyjaśnianie są względnie niezależne).

[26] W tym przypadku zwolennik ZKOS może jednak dość łatwo się bronić. Wystarczy, by stwierdził, że tylko zdolności związane z „centralnym" poznaniem – zwłaszcza przeprowadzanie rozumowań – mogą zostać w taki sposób wyjaśnione. (Por. jednak Matthews 2007: 36–38, 69–84, gdzie tego rodzaju stanowisko jest krytykowane na podstawie faktycznego stanu empirycznych badań nad przeprowadzaniem rozumowań).

przyjęcie, że mamy do czynienia z alternatywą międzyzałożeniem korespondencji a antynaturalistycznym separacjonizmem; alternatywą wyczerpującą wszelkie dostępne opcje teoretyczne. Gdyby taka ocena sytuacji była poprawna, to przyjęcie ZKOS co prawda skutkowałoby całkowitym pominięciem ważnych różnic między tym, co osobowe, a tym, co subosobowe, jednak byłaby to cena, którą należy rzekomo zapłacić za odrzucenie separacjonizmu i pozostanie na stanowisku naturalistycznym. ZKOS byłoby „jedyną opcją" dla naturalisty. Jednak tu trafiamy na kolejną zasadniczą słabość założenia korespondencji. Na jakiej bowiem podstawie sądzi się, że jest ono jedyną możliwą, przyjazną naturalizmowi interpretacją opozycji osobowe–subosobowe? ZKOS jest na tyle silne, na ile rzeczywiście nie ma dla niego naturalistycznej alternatywy. W następnej sekcji pokażę, że taka alternatywa istnieje. Mówiąc obrazowo, przewaga tej propozycji nad ZKOS polega (między innymi) na tym, iż pozwala nam ona „zjeść ciastko i mieć ciastko". Pozwala mianowicie pozostać na stanowisku naturalistycznym, nie ignorując jednocześnie filozoficznie istotnych różnic zachodzących między poziomem osobowym a subosobowym. Utrzymywanie ZKOS jako „jedynej sensownej opcji" dla naturalisty jest zupełnie nieuzasadnione.

Warto podsumować pokrótce ustalenia tej sekcji. Wskazałem cztery racje, dla których ZKOS może być uznawane za uzasadnione, a być może nawet jedyne teoretycznie „dopuszczalne" ujęcie relacji między poziomem osobowym a subosobowym. Pokazałem, że żadna z tych racji nie jest przekonująca. Założenie korespondencji między poziomem osobowym a subosobowym nie stanowi ani wyrazu pojęciowych zobowiązań psychologii potocznej, ani dobrej hipotezy empirycznej, ani konsekwencji wiarygodnych założeń dotyczących natury redukcji i wyjaśniania, ani wreszcie jedynego zgodnego z naturalizmem ujęcia relacji międzypoziomowych. Mówiąc zatem wprost: nie ma dobrych racji, by utrzymywać ZKOS. Pora zatem na zaprezentowanie alternatywnego ujęcia relacji między poziomem osobowym a subosobowym.

5.2. Poziom osobowy i subosobowy: interpretacja mechanistyczna

5.2.1. Poziom osobowy i subosobowy jako poziomy mechanizmów

W rozdziale 2 (podrozdział 2.2) zostało omówione zagadnienie poziomów organizacji mechanizmów i relacji międzypoziomowych w wyjaśnieniach mechanistycznych. Obecnie moim celem jest wykorzystanie tego sposobu myślenia o poziomach organizacji oraz związanych z nimi poziomach wyjaśniania w celu nadania naturalistycznej, a jednocześnie odchodzącej od założenia korespondencji, interpretacji opozycji osobowe–subosobowe. Bronioną tu propozycję można skrótowo zawrzeć w twierdzeniu: *poziom osobowy i subosobowy to poziomy wyjaśniania odpowiadające różnym poziomom mechanistycznej organizacji systemu poznawczego*. Pozostała część tego rozdziału będzie poświęcona rozwinięciu tej ogólnej idei oraz prześledzeniu jej konsekwencji[27].

Wydaje się, że mechanicyzm dostarcza dość jednoznaczną interpretację poziomu subosobowego. Zwróćmy bowiem uwagę na fakt, że wyjaśnianie odwołujące się do mechanizmów z niższego poziomu organizacji wykazuje w zasadzie wszystkie cechy charakterystyczne wyjaśniania subosobowego, które wymieniono w sekcji 5.1.1. Wyjaśnienia mechanistyczne pokazują, jak określone zjawiska – to znaczy zdolności poznawcze – są umożliwiane przez zorganizowane, działające komponenty mechanizmu. Wskazują one wewnętrzne, mechanistyczne *warunki umożliwiające* systemowi realizowanie określonych funkcji. Zauważmy też, że elementy składowe mechanizmów

[27] Należy od razu zaznaczyć, że pomysł, by interpretować dystynkcję osobowe–subosobowe przez pryzmat mechanicyzmu, nie jest całkowicie nowy. Mechanistyczne odczytanie tej dystynkcji proponowali już Bechtel i Abrahamsen (1993) oraz Herschbach (2008). Propozycja ta nie została jednak szerzej rozwinięta i nie przedostała się do głównego nurtu filozofii umysłu i kognitywistyki. Tutaj chcę zająć się przede wszystkim tym pierwszym brakiem, mam bowiem zamiar rozwinąć mechanistyczne odczytanie opozycji osobowe–subosobowe, szczególnie skupiając się na jego konsekwencjach dla zagadnienia relacji między projektem naturalizacji intencjonalności a problemem statusu eksplanacyjnego reprezentacji w kognitywistyce.

poznawczych występujące w roli eksplanansów będą zawsze strukturami o charakterze *subosobowym*: neuronalnymi, neurofizjologicznymi, obliczeniowymi czy neurooobliczeniowymi. Predykaty, za pomocą których opisujemy komponenty, działania i organizację mechanizmów poznawczych, nie mogą być orzekane o podmiotach czy systemach poznawczych jako całościach. Ze względu na te analogie można uznać, że subosobowe wyjaśnienie pewnego zjawiska poznawczego to inaczej wyjaśnienie tego zjawiska za pomocą odpowiedzialnego za nie mechanizmu. Wyjaśnienia subosobowe to wyjaśnienia odwołujące się do mechanizmów, na które składają się zorganizowane komponenty i działania z niższego poziomu organizacji systemu poznawczego. Poziom subosobowy to poziom wyjaśniania mechanistycznego.

Mechanistyczne odczytanie poziomu subosobowego niesie ze sobą dwie nowe konsekwencje, która każą nieco zmodyfikować przedstawioną w sekcji 5.1.1, wyjściową charakterystykę tego poziomu. Pierwsza modyfikacja wiąże się z faktem, że przy proponowanej tu interpretacji poziom subosobowy może zostać ujęty nie tylko jako epistemiczny poziom wyjaśniania, ale również jako poziom w sensie metafizycznym. Z jednej strony jest to poziom w sensie epistemicznym (poziom wyjaśniania), na którym to wyjaśniamy zjawiska przez odwołanie do wewnętrznej, mechanistycznej architektury systemu poznawczego. Z drugiej – to metafizyczny poziom organizacji, na którym znajdują się elementy składowe wewnętrznych mechanizmów poznawczych. Mówiąc o subosobowych „strukturach", „stanach" czy „komponentach", ma się w takiej perspektywie na myśli właśnie elementy składowe wewnętrznych mechanizmów poznania.

Druga wynikająca z mechanicyzmu modyfikacja rozumienia poziomu subosobowego ma z kolei związek z faktem, iż kompozycyjna organizacja mechanizmów jest hierarchiczna. Działanie komponentu mechanizmu A może zostać wyjaśnione za pomocą mechanizmu B, zawierającego komponent, którego działanie jest wyjaśniane przez mechanizm C z jeszcze niższego poziomu organizacji i tak dalej. Otóż sądzę, że każdy poziom organizacji systemu poznawczego, na którym znajdują się mechanizmy eksplanacyjnie istotne przy wyjaśnianiu zjawisk poznawczych, powinien być sklasyfikowany jako

subosobowy. W związku z tym zamiast mówić o jednym poziomie subosobowym, powinniśmy mówić o całej „kaskadzie" *poziomów* subosobowych, znajdujących się na coraz niższych „piętrach" hierarchicznej organizacji systemu poznawczego.

Kiedy jednak próbujemy nadać mechanistyczną interpretację wyjaśnieniom z poziomu osobowego, pojawiają się trudności. Wyjaśnienia z tego poziomu nie przypominają bowiem mechanistycznych. W jaki sposób wpisują się zatem w mechanistyczną wizję organizacji i wyjaśniania systemu poznawczego? Pomocne będzie tu rozróżnienie Bermúdeza (2005: 31–35) na wyjaśnienia wertykalne i horyzontalne. Wyjaśnienie wertykalne to inaczej *międzypoziomowe* wyjaśnienie zjawiska z wyższego za pomocą faktów z niższego poziomu. Powstaje tu rzecz jasna problem, co dokładnie ma się na myśli, mówiąc o „poziomach". Jednak w kontekście prowadzonych rozważań można uznać, że chodzi tu o poziomy mechanistycznej organizacji pewnego systemu. Wyjaśnienia subosobowe są zatem wertykalne w tym sensie, że wyjaśniają zjawiska z wyższego poziomu kompozycyjnej organizacji systemu poznawczego za pomocą struktur z niższego poziomu tej organizacji.

Bermúdez twierdzi zarazem, że wyjaśnienia osobowe mają charakter *horyzontalny*, czyli są wewnątrzpoziomowe. Poruszając się na poziomie osobowym, wyjaśniamy dane zjawisko przez odwołanie się do innych zjawisk z tego samego poziomu. Dokładniej rzecz biorąc, w wyjaśnieniach osobowych: (1) eksplananse znajdują się na tym poziomie, co eksplananda; (2) wystąpienie eksplanansu poprzedza czasowo wystąpienie eksplanandum; (3) eksplanans jest przyczynowo odpowiedzialny za pojawienie się eksplanandum. Z perspektywy proponowanej przez Bermúdeza wyjaśnienia z poziomu osobowego to zatem *wewnątrzpoziomowe, etiologiczne wyjaśnienia przyczynowe* (por. Craver 2007: 74, 93–104). Wyjaśniając czyjeś działanie za pomocą przekonania czy pragnienia tej osoby, nie „schodzimy" na niższy poziom, lecz wskazujemy przyczyny tego działania, które je poprzedzają temporalnie i które są zlokalizowane na tym samym poziomie, co samo działanie.

Tabela 3. Zestawienie osobowego i subosobowego poziomu wyjaśniania przy mechanistycznej interpretacji dystynkcji osobowe–subosobowe

Poziom wyjaśniania	Poziom osobowy	Poziom subosobowy
Rodzaj wyjaśniania: horyzontalne versus wertykalne	Wyjaśnianie horyzontalne (wewnątrzpoziomowe) – eksplanans i eksplanandum znajdują się na tym samym poziomie organizacji systemu poznawczego.	Wyjaśnianie wertykalne (międzypoziomowe) – eksplanans znajduje się na niższym poziomie organizacji systemu poznawczego niż eksplanandum.
Poziom organizacji systemu poznawczego, na którym jest sformułowane wyjaśnienie	Poziom systemowy, czyli system poznawczy (organizm) jako całość zaangażowana w interakcje ze środowiskiem.	Wewnętrzne mechanizmy systemu poznawczego, ich zorganizowane, działające komponenty.
Rodzaj pytania, na które odpowiada wyjaśnienie	Pytanie o to, dlaczego zaszło eksplanandum.	Pytanie o to, jak eksplanandum jest umożliwiane przez stojący u jego podstaw mechanizm.
Natura eksplanandum	Poszczególne działania wykonywane przez system poznawczy (organizm).	Zdolności (dysfunkcje) systemu poznawczego lub jego komponentów.
Natura eksplanansu	Postawy propozycjonalne rozumiane jako własności systemowe (własności wyższego rzędu).	Wewnętrzna, mechanistyczna architektura systemu poznawczego (elementy składowe mechanizmów).

Akceptuję twierdzenie Bermúdeza, że wyjaśnienia osobowe są horyzontalne. Tym samym nie mogą być uznane za formę wyjaśnień mechanistycznych (które są z konieczności wertykalne). Chcę tu jednak konsekwentnie zachować mechanistyczne rozumienie poziomów. Wyjaśnienia osobowe to horyzontalne wyjaśnienia przyczynowe w obrębie *jednego mechanistycznego poziomu organizacji* systemu poznawczego. Na którym z poziomów tej organizacji jest

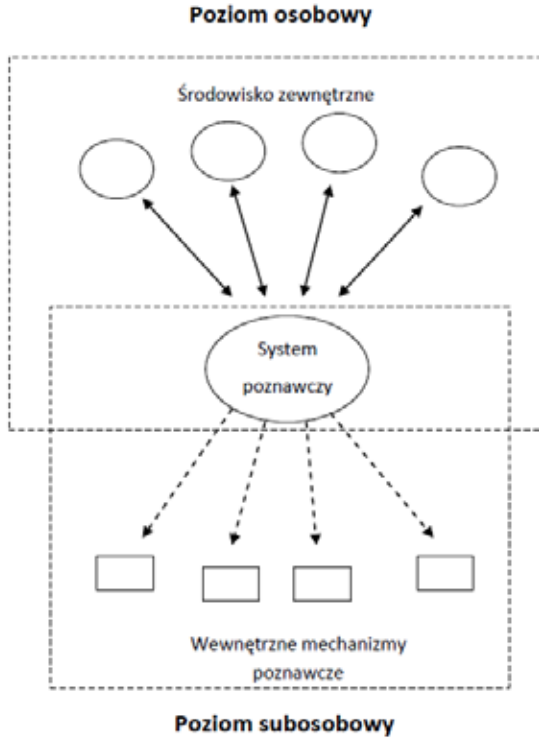
jednak aplikowana osobowa aparatura pojęciowa psychologii potocznej? Otóż postuluję, że osobowemu poziomowi wyjaśniania odpowiada poziom systemu poznawczego (organizmu) jako całości zaangażowanej w wielorakie interakcje z zamieszkiwanym przez siebie środowiskiem. Wyjaśnienia osobowe nie zaglądają „do wewnątrz" systemu, aby odpowiedzieć na pytanie o to, jak on działa (tu i dalej por. tabela 3). Nie skupiają się one na interreagujących komponentach systemu. Koncentrują się raczej na poziomie tego systemu jako całości i sytuują ten system w ramach jego środowiska. Dostarczają zatem odpowiedzi na pytania o to, dlaczego system poznawczy działa w określonych okolicznościach w taki, a nie inny sposób. W roli eksplanansu nie występują elementy wewnętrznej mechanistycznej architektury systemu (struktury subosobowe), lecz przyczynowo efektywne *własności systemowe*, których podmiotem jest system poznawczy czy organizm pojęty jako całość.

W świetle bronionej tu propozycji zachodzi zasadnicza różnica między rolą eksplanacyjną osobowych stanów intencjonalnych a rolą eksplanacyjną stanów i struktur o charakterze subosobowym. Atrybucja przekonań i pragnień nie daje nam wglądu w to, *jak* działa system poznawczy. Przypisując innym przekonania, pragnienia, intencje czy oczekiwania, nie opisujemy wewnętrznych mechanizmów poznania. Postawy propozycjonalne mają raczej wyjaśniać, dlaczego system poznawczy działa w taki, a nie inny sposób.

Rzecz jasna wskazanie, jaką rolę eksplanacyjną odgrywają postawy propozycjonalne, nie rozwiązuje jeszcze problemu, czym one są. Z perspektywy mechanistycznej przekonania i pragnienia powinny zostać uznane za własności wyższego rzędu, czyli własności systemowe. „Bycie przekonanym, że *p*", „pragnienie, by *p*" czy „oczekiwanie, że *p*" to własności egzemplifikowane przez systemy poznawcze, a nie przez ich wewnętrzne, subosobowe komponenty. Jakiego rodzaju własnościami systemowymi są jednak przekonania, pragnienia, nadzieje, oczekiwania i inne postawy propozycjonalne? Zgodnie z bronionym tu podejściem, jak to ujmuje Eric Schwitzgebel, „posiadać postawę to żyć w określony sposób" (2013: 76). Systemy poznawcze stanowią podmioty przekonań i pragnień ze względu na ich dyspozycje do działania w środowisku według określonych wzorców

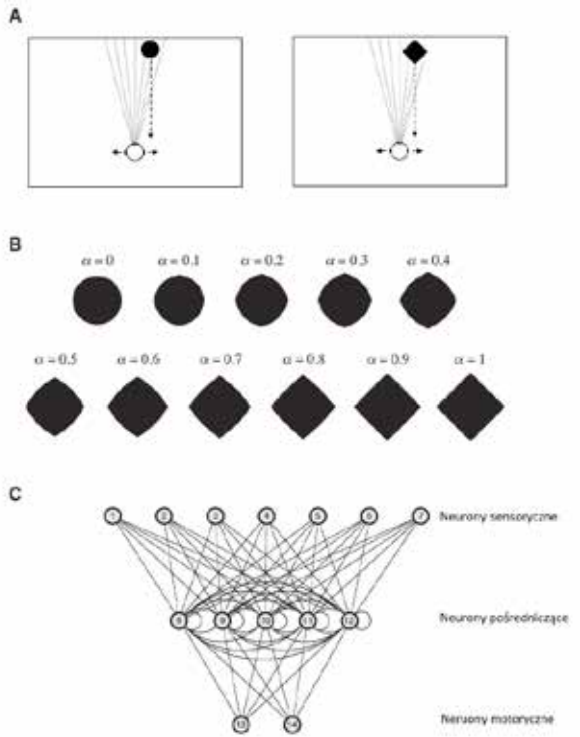
i regularności. Postawy propozycjonalne to *własności dyspozycyjne* systemów poznawczych (por.: Bechtel, Abrahamsen 1993; Clark 1993: 189–219; Audi 1994; Baker 1995: 153–192; Matthews 2007: 123–256; Ratcliffe 2007: 205–211; Schwitzgebel 2002, 2013). Pragnąć, by *p*, to wykazywać zestaw dyspozycji do myślenia, działania i odczuwania w sposób charakterystyczny dla posiadaczy pragnienia, by *p*; to znaczy (między innymi) w sposób prowadzący do zrealizowania tego, że *p*. Posiadać przekonanie, że *p*, to mieć dyspozycje do myślenia, działania i odczuwania w sposób charakterystyczny dla posiadaczy przekonania, że *p*; to znaczy (między innymi) w sposób, który bierze pod uwagę to, że *p*, przy planowaniu działań, podejmowaniu decyzji albo odpowiadaniu na czyjeś pytania o to, czy *p*. Tego rodzaju dyspozycyjne podejście do postaw propozycjonalnych rodzi z pewnością cały szereg filozoficznych pytań i wątpliwości, dlatego poświęcę mu oddzielny podrozdział (5.3).

Zgodnie z moją propozycją psychologia potoczna koncentruje się nie tyle na wewnętrznej strukturze systemu poznawczego, co raczej na interakcjach tego systemu z zamieszkiwanym przez niego środowiskiem zewnętrznym (por. rysunek 10). Mówiąc o środowisku „zamieszkiwanym" przez system, chcę zwrócić uwagę na fakt, że wyjaśnienia na poziomie osobowym koncentrują się nie na dowolnych elementach otoczenia fizycznego, które jakoś oddziałują przyczynowo na system, lecz na tych elementach, które są *pragmatycznie* czy *funkcjonalnie* relewantne dla tego systemu; takich elementach, względem których system ten musi „ustosunkowywać" się poznawczo i behawioralnie. W przypadku istot ludzkich środowisko w takim sensie obejmuje także otoczenie społeczne i kulturowe. Funkcja epistemiczna psychologii potocznej polega przede wszystkim na (1) określaniu wzorców prawdopodobnych działań organizmu zorientowanych na stany rzeczy (własności, obiekty, procesy) w jego środowisku, a także (2) wskazywaniu elementów tego ostatniego, na które system jest poznawczo i behawioralnie „czuły" (por. Bechtel, Abrahamsen 1993). Psychologia potoczna pozwala na zrozumienie, jak system poznawczy jest „podłączony" do swojego świata.



Rysunek 10. Poziom osobowy i subosobowy – interpretacja mechanistyczna. Opisy i wyjaśnienia z poziomu osobowego sytuują system poznawczy jako całość w jego środowisku. Podwójne ciągłe strzałki symbolizują wzory interakcji między systemem a elementami jego otoczenia. Opisy i wyjaśnienia z poziomu subosobowego koncentrują się na wewnętrznych mechanizmach, które stoją u podstaw zdolności cechujących system poznawczy. Przerywane strzałki symbolizują kompozycyjną relację między systemem a elementami składowymi (zorganizowanymi, działającymi komponentami) jego wewnętrznych mechanizmów

Spróbuję teraz wykorzystać prosty, poglądowy przykład, aby nadać bardziej konkretny kształt przedstawionemu do tej pory szkicowi teoretycznemu. Weźmy pod uwagę sztuczny, wirtualny organizm



Rysunek 11. Wirtualny organizm opisany przez Randalla Beera. A. Organizm zajmuje się „łapaniem" obiektów w kształcie okręgu oraz unikaniem obiektów przybierających kształt rombu. B. Rodzaje obiektów, na które jest eksponowany organizm. C. Układ połączeń w sieci neuronowej kontrolującej ruch organizmu. Źródło: Beer 2003: 213

stworzony i opisany kompleksowo przez Randalla Beera (2003; por. także de Pinedo, Noble 2008). Ten prosty system zamieszkuje nieskomplikowane wirtualne środowisko. Całe jego „życie" sprowadza się do odpowiedniego reagowania na wertykalnie przemieszczające się, „spadające" w jego kierunku obiekty. Organizm ten jest w stanie wykonywać horyzontalne ruchy pozwalające mu na „łapanie" jednych obiektów oraz unikanie innych (por. rysunek 11-A). Jego za-

danie polega na chwytaniu obiektów o kształcie okręgu i unikaniu obiektów przyjmujących kształt rombu. Na organizm spadają jednak nie tylko okręgi i romby w „czystej" postaci, ale także obiekty stanowiące, w różnych proporcjach (określonych za pomocą zmiennej α), hybrydy tych kształtów (por. rysunek 11-B).

Za ruch wirtualnego organizmu odpowiada trójwarstwowa sieć neuronowa podłączona do wirtualnego „oka" emitującego siedem promieni (por. rysunek 11-C). Sygnały z oka aktywują pierwszą, sensoryczną warstwę sieci (składającą się z siedmiu neuronów), która następnie pobudza warstwę pośredniczącą (pięć neuronów), a ta pobudza z kolei warstwę motoryczną (dwa neurony). Neurony w warstwie motorycznej są podłączone do wirtualnych efektorów organizmu, dzięki czemu sterują one jego ruchem. Neurony w warstwie pośredniczącej są powiązane nie tylko z tymi znajdującymi się w dwóch pozostałych warstwach za pomocą połączeń wyprzedzających, ale też wzajemnie za pomocą połączeń zwrotnych.

Co istotne, układ wag połączeń synaptycznych między neuronami opisywanej sieci stanowi wynik działania algorytmu genetycznego. Innymi słowy – sieć, dzięki której porusza się prosty organizm opisywany przez Beera, stanowi efekt wielu pokoleń wirtualnej ewolucji przez dobór naturalny, gdzie kolejne wirtualne pokolenia były eksponowane na spadające okręgi i romby. Poziom dostosowania osobników zależał od zdolności do percepcyjnego odróżniania jednych obiektów od innych zgodnie z ich przynależnością kategorialną, co umożliwiało „łapanie" okręgów oraz unikanie rombów. Organizm analizowany przez Beera to „potomek" swoich najlepiej dostosowanych przodków. Jednak zamieszkuje on środowisko bardziej złożone od tego, w którym ewoluowali jego antenaci. Jak bowiem wcześniej wspomniałem, musi on percepcyjnie rozróżniać i behawioralnie reagować nie tylko na okręgi i romby, ale także na obiekty o niepewnej przynależności kategorialnej, stanowiące amalgamaty tych kształtów.

Beer zmierza do sformułowania bardzo drobiazgowego i precyzyjnego wyjaśnienia tego, jak sieć neuronowa umożliwia określone wzorce interakcji między organizmem a jego środowiskiem. Mówiąc zaś bardziej precyzyjnie, badacz ten chce wyjaśnić, w jaki

sposób wirtualnemu organizmowi udaje się dysponować zdolnością do *aktywnej percepcji kategorialnej* zmierzających w jego kierunku obiektów[28]. Na czym polega i w czym objawia się owa zdolność? Kiedy przyglądamy się zachowaniom organizmu, okazuje się, że wykazuje on tendencję do aktywnego percepcyjnego „skanowania" obiektów i reagowania na nie zgodnie z ich przynależnością kategorialną. Mamy tu do czynienia z trzyetapowym procesem. Kiedy w kierunku organizmu zaczyna zmierzać obiekt, organizm ten najpierw skupia się na nim – to znaczy zajmuje położenie, dzięki któremu siatkówka jego wirtualnego oka jest scentrowana na percypowanym obiekcie. W następnej fazie organizm wykonuje serię eksploracyjnych ruchów tam i z powrotem, które Beer (2003: 215) nazywa „aktywnym skanowaniem" obiektu. W trzeciej fazie dochodzi do podjęcia decyzji: organizm pozostaje bezpośrednio pod opadającym obiektem lub unika go przez wykonanie ruchu w jedną ze stron (Beer pokazuje, że można nawet wyznaczyć moment, w którym organizm definitywnie „zobowiązał" się do określonej reakcji behawioralnej i stał się poznawczo nieczuły na ewentualne zmiany kształtu obiektu).

Należy zauważyć, że wzorzec behawioralnych reakcji organizmu na różne hybrydowe obiekty wykazuje własności charakterystyczne dla percepcji kategorialnej (Beer 2003). Cechy te to: etykietowanie (*labeling*) oraz dyskryminacja (*discrimination*). Etykietowanie polega na przyporządkowywaniu bodźców stanowiących kontinuum do odrębnych, dyskretnych kategorii. W przypadku systemu opisywanego przez Beera etykietowanie przejawia się w tym, iż istnieje ściśle określona granica oddzielająca reakcje polegające na „łapaniu" od reakcji polegających na unikaniu spadającego obiektu. Granica ta jest wyznaczana (w przybliżeniu) przez obiekty, dla których wartość α wynosi 0.3. Można więc powiedzieć, że dla wirtualnego organizmu obiekty o $\alpha<0.3$ wyznaczają kategorię odrębną (są one kategoryzowane jako okręgi i „chwytane") względem kategorii, do której należą pozostałe obiekty (kategoryzowane jako romby i unikane). Z kolei

[28] Trzeba zaznaczyć, że to właśnie posiadanie zdolności do aktywnej percepcji kategorialnej czyni według Beera ten wirtualny organizm systemem minimalnie poznawczym (zdolnym do minimalnie poznawczego zachowania).

dyskryminacja jako własność percepcji kategorialnej polega na tym, że bodźce o różnej przynależności kategorialnej są lepiej odróżniane od siebie, niż bodźce należące do tej samej kategorii. W przypadku wirtualnego organizmu Beera dyskryminacja percepcyjna przejawia się w tym, że organizm najlepiej rozróżnia od siebie nie obiekty znajdujące się wewnątrz jednej kategorii, lecz te, dla których wartość α wynosi około 0.3, to znaczy takie, które są najbardziej zbliżone do międzykategorialnej granicy.

Omawiany artykuł jest w dużym stopniu poświęcony dostarczeniu matematycznego opisu sieci neuronowej kontrolującej wirtualny organizm, a także opisaniu interakcji między tą siecią, organizmem i jego środowiskiem. Opis ten wykorzystuje aparaturę pojęciową teorii systemów dynamicznych. Dzięki dysponowaniu nim jesteśmy w stanie zrozumieć, jak wirtualnemu organizmowi udaje się percepcyjnie odróżniać obiekty zgodnie z ich przynależnością kategorialną. Zauważmy, że Beer jest w istocie zainteresowany dostarczeniem wyjaśnienia mechanistycznego. Wyjaśniając zdolność wirtualnego organizmu do kategoryzacji percepcyjnej, odwołuje się on do wewnętrznej, mechanistycznej struktury, czyli sieci neuronowej, na podstawie której aktywności organizm ten wykazuje określone wzorce zachowania. To funkcjonowanie tej sieci umożliwia organizmowi wchodzenie w określone rodzaje interakcji z okręgami, rombami i obiektami hybrydowymi.

Jedna z najważniejszych konsekwencji, jakie Beer wyprowadza ze swojej analizy, ma jednoznacznie antyreprezentacjonistyczny wydźwięk. Jak wyraźnie zaznacza ten autor, wewnętrzna maszyneria sterująca wirtualnym organizmem nie wykorzystuje nic, co choćby przypomina (funkcjonalnie) reprezentacje. Organizm nie tylko nie dysponuje żadnym modelem okręgów i rombów. Nie ma w nim nawet żadnych struktur, które mogłyby „kandydować" do miana detektorów czy receptorów tych dwóch kategorii obiektów. Okazuje się bowiem, że „okręgi" i „romby" nie mają żadnych wewnętrznych korelatów w sieci neuronowej sterującej wirtualnym organizmem, niczego, co systematycznie współwystępowałoby wyłącznie z pierwszą, ale nie drugą kategorią obiektów lub *vice versa*. Co więcej, Beer zwraca uwagę na fakt, że różne intencjonalne kategorie, za pomocą

których możemy opisywać wirtualny organizm, nie mają żadnych wewnętrznych, neuronalnych odpowiedników. Na przykład w opisywanym przez Beera systemie poznawczym nie istnieje żadna wewnętrzna struktura czy wzór pobudzenia neuronów, który można byłoby utożsamić z „preferencją" do unikania rąbów albo „decyzją", aby uniknąć rombu (nie ma też nic, co byłoby choćby systematycznie skorelowane z taką preferencją czy decyzją).

Opisane wyżej konstatacje można podsumować za pomocą twierdzenia, iż wewnątrz wirtualnej neuronalnej architektury organizmu opisywanego przez Beera nie istnieje nic, co można by użytecznie opisać w kategoriach reprezentacyjnych. Mechanistyczne wyjaśnienie zdolności do kategoryzacji percepcyjnej w tym przypadku obywa się całkowicie bez pojęcia reprezentacji. Nie ma powodów, aby się nie zgodzić z taką antyreprezentacjonistyczną diagnozą Beera. Warto jednak spojrzeć na tę konkluzję przez pryzmat mechanistycznego odczytania dystynkcji osobowe–subosobowe. Jak sądzę, Beer może twierdzić jedynie, że *subosobowa* architektura opisywanego przez niego minimalnie poznawczego systemu nie wykorzystuje wewnętrznych, *subosobowych* reprezentacji. Nie potrzebujemy pojęcia wewnętrznych reprezentacji do wyjaśnienia tego, *jak* wirtualnemu organizmowi udaje się kategoryzować percepcyjnie wirtualne obiekty[29]. Spróbujmy jednak zapytać o to, jak w omawianym tu

[29] Zwróćmy jednak uwagę, że sytuacja ta mogłaby się zmienić. Możemy sobie wyobrazić kolejne stadia ewolucji naszego wirtualnego organizmu – czy raczej populacji organizmów, do której on należy. Być może doszłoby do wzrostu złożoności środowiska. Na przykład romby i okręgi (oraz ich hybrydy) mogłyby w pewnym momencie przyspieszyć swoje spadanie w kierunku organizmów. Być może zaczęłyby też pojawiać się sytuacje, w których na pojedynczy organizm spada nie jeden, lecz dwa obiekty na raz. Możemy spekulować, że tego rodzaju zmiany wytworzyłyby presje selekcyjne napędzające rozwój złożoności behawioralnej jednostek w populacji. Rozwojowi temu rzecz jasna musiałyby towarzyszyć zmiany w strukturze i układzie wag sieci neuronowej sterującej zachowaniami jednostek. Niewykluczone, że w wyniku „rozrostu" sieci pojawiłby się problem związany z opóźnieniem między pojawianiem się obiektu/obiektów a wygenerowaniem odpowiedzi behawioralnej. Toteż w wyniku wzrostu złożoności wewnętrznej sieci neuronowej (oraz związanego z tym wydłużenia interwału czasowego między pojawieniem się obiektu a wygenerowaniem przez sieć stosownej „odpowiedzi" behawioralnej) organizmy nie byłyby już w stanie

przypadku rzeczy mają się na poziomie *osobowym*. Co potrafi ten *system jako całość*? Czy może on jako całość egzemplifikować jakieś własności intencjonalne? Jakie to (ewentualnie) własności?

Beer ma z pewnością rację, twierdząc, że nie istnieją żadne wewnętrzne odpowiedniki przypisywanych organizmowi kompetencji i stanów intencjonalnych, takich jak przykładowo aktywna percepcja obiektów;rozpoznawanie okręgów i rombów; kategoryzowanie percepcyjne obiektów; podejmowanie decyzji o podjęciu jakiegoś działania; okazjonalna niepoprawna kategoryzacja i zła decyzja (na przykład skutkująca „złapaniem" rombu); preferencja (czy proto-pragnienie) do „łapania" okręgów; preferencja do unikania rombów. Swoisty paradoks polega jednak na tym, że Beer w praktyce *rzeczywiście opisuje* analizowany organizm za pomocą wymienionych kategorii (de Pinedo, Noble 2008). Co prawda nie istnieje coś takiego, jak wewnętrzny, neuronalny korelat decyzji, aby uniknąć rombu, jednak nie zmienia to faktu, że organizm jest opisywany jako podejmujący decyzje. Beer nieustannie przypisuje temu prostemu systemowi zdolności i stany intencjonalne (percepcja *czegoś*, kategoryzacja/dyskryminacja *czegoś*, podejmowanie decyzji *o czymś*, preferowanie *czegoś*). Co więcej, jak wskazują Manuel de Pinedo i Jason Noble (2008), atrybucja tego rodzaju stanów i zdolności jest Beerowi teoretycznie niezbędna. Co stanowi wszakże dla tego autora przedmiot wyjaśniania, czyli eksplanandum? Jest to zdolność do kategoryzacji percepcyjnej i podejmowania na tej podstawie trafnych decyzji behawioralnych. Złożony matematyczny opis dostarczany przez Beera byłby całkowicie niezrozumiały, gdybyśmy nie wiedzieli, co ma

behawioralnie „nadążać" za swoim środowiskiem. Możemy sobie wyobrazić, że taki obrót spraw doprowadziłby do wyłonienia się strategii kompensującej to opóźnienie, a polegającej na wygenerowaniu wewnętrznego *modelu* środowiska (por. Grush 1997). Model ten odzwierciedlałby statystyczne prawidłowości środowiska i pozwalał na przewidzenie kształtu oraz liczby kolejnych spadających obiektów, jeszcze zanim do systemu dotrą stosowne informacje sensoryczne. (Zakładam, że takie prawidłowości w środowisku zachodzą i że proces generowania obiektów o określonych kształtach nie jest całkowicie losowy). Gdyby ten spekulatywny scenariusz się ziścił, bylibyśmy świadkami sytuacji, w której wirtualna ewolucja przez dobór naturalny doprowadziła do powstania *reprezentacyjnych* systemów poznawczych (por. Godfrey-Smith 2002).

on wyjaśniać. Jak to ujmują de Pinedo i Noble, „spoglądając jedynie na poziom mechanizmów, byłoby bardzo trudne, a może nawet niemożliwe dostrzeżenie, że cała ta złożoność służy łapaniu okręgów i unikaniu rombów" (2008: 96).

Zgadzam się z de Pinedo i Noblem (2008), że ów paradoks – wynikający z napięcia zachodzącego między opisem systemu stworzonego przez Beera w kategoriach intencjonalnych a nieistnieniem wewnętrznych, subosobowych odpowiedników tych kategorii – jest jedynie pozorny. Wspomniane napięcie wynika z nieuzasadnionego oczekiwania, jakoby tego rodzaju odpowiedniość musiała zachodzić, aby opis intencjonalny był uzasadniony czy prawomocny. Tymczasem aktywna percepcja, rozpoznawanie, kategoryzowanie, podejmowanie decyzji i posiadanie preferencji to zjawiska, które znajdujemy tylko na systemowym poziomie organizacji. Poprawność atrybucji każdej z tych zdolności czy każdego stanu intencjonalnego nie opiera się na wewnętrznej budowie wirtualnego organizmu, lecz na *wzorcach jego interakcji z zamieszkiwanym przez niego środowiskiem* (por. Dennett 2008). Skąd wszakże wiemy, że organizm wykorzystuje zdolność kategoryzacji percepcyjnej w celu „łapania" okręgów i unikania rombów? Skąd wiemy, że preferuje on okręgi, a unika rombów? Jak się wydaje, przypisujemy organizmowi tego rodzaju zdolności czy stany dzięki przyjęciu perspektywy *ekologicznej*, „lokującej" go w ramach jego środowiska. Przyjmując ją, spoglądamy na „dyskryminacyjne", oparte na percepcji działania organizmu względem obiektów, na które natrafia on jako całość w swoim środowisku. Również treści, które można przypisywać stanom intencjonalnym organizmu opisanego przez Beera, są ściśle związane z tym, jak percepcyjnie dyskryminuje on obiekty i jak na nie praktycznie reaguje. Mówimy, że percypuje on „okręgi" i „romby", ponieważ właśnie te dwie kategorie obiektów są dla niego pragmatycznie istotne, jest na nie poznawczo i behawioralnie „czuły" (powtórzmy: wewnątrz organizmu nie ma nic, co byłoby chociażby systematycznie skorelowane/współzmienne z tymi kategoriami). W momencie *t* organizm *podjął decyzję* dotyczącą „złapania" spadającego obiektu nie dlatego, że w *t* aktywowała się jakaś wewnętrzna struktura kodująca treść „należy to złapać", lecz dlatego, że w momencie tym organizm be-

hawioralnie „zobowiązał" się do określonej reakcji (począwszy od *t*, nie reaguje on już na ewentualne zmiany kształtu obiektu). We wszystkich wymienionych przypadkach poruszamy się na poziomie osobowym (systemowym), a nie subosobowym (subsystemowym)[30].

Powyżej koncentrowałem się na naturze opisu wirtualnego organizmu w osobowych kategoriach intencjonalnych. Warto jednak postawić pytanie o to, czy moglibyśmy *wyjaśnić* jego działanie za pomocą tego rodzaju kategorii. Wydaje się, że odpowiedź powinna być twierdząca, jednak musimy brać poprawkę na specyfikę wyjaśnienia, o którym mówimy. Zgodnie z proponowanym tu ujęciem na poziomie osobowym nie sformułujemy *mechanistycznego* wyjaśnienia dającego odpowiedź na pytanie: „Jak wirtualnemu organizmowi udaje się dyskryminować obiekty zgodnie z ich przynależnością kategorialną?". Wyobraźmy sobie jednak, że ktoś po raz pierwszy obserwuje zachowanie wirtualnego organizmu, widzi, jak „ucieka" on od rombu, i pyta nas: „Dlaczego organizm uniknął kontaktu ze spadającą figurą?". Jak brzmiałaby odpowiedź? Moglibyśmy powiedzieć, że organizm ten posiada zdolność do *rozpoznawania* percepcyjnego, do której z dwóch kategorii – okręgów czy rombów – należy dany obiekt. Posiada on także *preferencję* do unikania obiektów rozpozna-

[30] Jak widać, w przyjmowanym tu rozumieniu termin „osobowy" może być używany zamiennie z „systemowy". W szczególności, mówiąc o „poziomie osobowym", nie mam zamiaru odnosić się wyłącznie do „poziomu" *bytów osobowych*, czyli posiadających takie własności, jak podmiotowość moralna, wolna wola, samostanowienie czy poczucie tożsamości osobowej. Dystynkcję osobowe–subosobowe można przeprowadzić dla *dowolnego* systemu poznawczego (nawet niezwykle prostego), który możemy opisywać i wyjaśniać, koncentrując się na nim jako całości (poziom osobowy, czyli systemowy i ekologiczny), jak również koncentrując się na organizacji i częściach składowych jego wewnętrznych mechanizmów (poziom subosobowy, czyli subsystemowy i mechanistyczny). Jak to ujmuje McDowell, każdy system poznawczy zdolny do „kompetentnego zamieszkiwania swojego środowiska" (1994: 196) możemy rozpatrywać na tak rozumianym poziomie osobowym. (Trzeba też zaznaczyć, że interpretacja dystynkcji „osobowe/subosobowe" za pomocą innej dystynkcji „systemowe/ /subsystemowe" jest zrelatywizowana do klasy systemów poznawczych. Nie ma rzecz jasna żadnego sensu aplikowanie jej do systemów niepoznawczych, takich jak systemy planetarne, układy krwionośne, pojedyncze neurony, samochody i tak dalej).

wanych jako romby. W tym przypadku (poprawnie) *skategoryzował* on spadającą figurę jako romb i właśnie dlatego podjął behawioralną *decyzję* o wykonaniu uniku. Odpowiadanie na dotyczące konkretnego działania pytanie „dlaczego?" – w przeciwieństwie do odpowiadania na pytanie „jak?", dotyczącego zdolności poznawczej – naturalnie sprzyja sformułowaniu właśnie tego rodzaju wyjaśnienia. Takie horyzontalne, osobowe wyjaśnienie odwołuje się do zestawu kategorii intencjonalnych (percepcja, preferencja, kategoryzacja, decyzja). Pozostaje ono poprawne nie ze względu na wewnętrzną architekturę systemu poznawczego, którego działanie jest wyjaśniane, lecz ze względu na to, że trafnie oddaje ono sposób, w jaki system ten, jako całość, zamieszkuje swoje środowisko.

Ostateczna diagnoza w sprawie statusu wirtualnego organizmu opisanego przez Beera jest zatem następująca. Nie dysponuje on reprezentacjami subosobowymi. Mechanistyczne wyjaśnienie jego zdolności obywa się bez pojęcia reprezentacji. Jednak na poziomie *osobowym* – na którym ujmujemy ten organizm jako całość zaangażowaną w różnorakie interakcje ze swoim środowiskiem – wydaje się on jednak systemem posiadającym proste, dość minimalne kompetencje i stany intencjonalne.

5.2.2. Interludium: McDowell o poziomie osobowym i wyjaśnieniach konstytutywnych

Zanim przejdę do omówienia konsekwencji takiej mechanistycznej koncepcji relacji osobowe–subosobowe, spróbuję naświetlić ją od jeszcze innej strony. Do bardzo zbliżonych konkluzji na temat relacji międzypoziomowych dochodzi John McDowell (1994). Także ten autor, rozwijając swoje stanowisko, omawia przykład stosunkowo prostego systemu poznawczego, tym razem jednak stanowiącego wynik realnego, a nie wirtualnego, doboru naturalnego. Nawiązując do artykułu *Co żabie oko mówi żabiemu mózgowi?* Jerome'a Lettvina i współpracowników (1959), McDowell podejmuje kwestię zdolności żab do percepcyjnej detekcji owadów. Kiedy zastanawiamy się nad naturą tej żabiej zdolności, powinniśmy wedle wspomnianego autora odróżnić dwa pytania. Pierwsze dotyczy kwestii, jak sys-

tem percepcyjny żaby umożliwia jej tego rodzaju detekcję. Według McDowella (1994) udzielenie odpowiedzi będzie wymagało pokazania, w jaki sposób części czy komponenty żaby (jej mózgu) umożliwiają temu organizmowi percepcyjne rozpoznawanie owadów. Jest to zatem pytanie o mechanistyczne warunki umożliwiające żabie posiadanie określonej funkcji poznawczej. Warunki te mają w ujęciu McDowella charakter subosobowy, natomiast zjawisko, które umożliwiają, czyli percepcja przedmiotów w otoczeniu – osobowy. Możemy powiedzieć, że badając subosobowe warunki umożliwiające fenomen z poziomu osobowego, poszukujemy mechanistycznego wyjaśnienia tego fenomenu.

Drugie wyróżnione przez McDowella (1994) zagadnienie to pytanie o to, co konstytuuje takie zjawisko z poziomu osobowego, jak percypowanie owadów. Kiedy w ten sposób postawimy sprawę, poszukujemy *konstytutywnego wyjaśnienia* percepcji. Wyjaśnienie tego rodzaju musi dostarczyć odpowiedzi na pytania: czym jest albo na czym polega percepcyjne rozpoznawanie przedmiotów przez żabę? Jakiego rodzaju stanem będzie percepcja tego, że w otoczeniu pojawił się owad czy też pożywienie? Według McDowella odpowiadanie na te pytania na podstawie faktów dotyczących *wewnętrznej* budowy żaby stanowi czynność zasadniczo chybioną i jałową poznawczo. Autor ten twierdzi, że poszukiwanie jakiegoś subosobowego komponentu żaby kodującego treść „to jest owad" – a przez to determinującego treść stanu percepcyjnego przypisywanego *żabie* – wynika z braku odpowiedniego rozróżnienia między poziomem osobowym („żabim") a subosobowym („sub-żabim"). Percypowanie obiektów jest bowiem umożliwiane, ale nie *konstytuowane* przez fakty z poziomu subosobowego. Nic wewnątrz żaby nigdy nie zobaczyło czy rozpoznało owada – tak samo jak nic wewnątrz organizmu opisanego przez Beera nigdy nie zobaczyło czy rozpoznało okręgu. Widzenie owada to bowiem własność wyróżniana na poziomie osobowym, który McDowell traktuje – podobnie jak przyjmuję w tej książce – jako poziom systemowy. To żaba jako całość, a nie żadna jej część, percypuje obiekty w swoim otoczeniu. To ona jest epistemicznie „wyczulona" na pojawianie się owadów i gotowa do odpowiedniej reakcji. Treści, które przypisujemy stanom percepcyjnym

żaby, są „widoczne" tylko z ekologicznego punktu widzenia, sytuującego zwierzę w ramach środowiska, z którym wchodzi ona w praktyczne interakcje i od którego jest praktycznie zależna. Owady pojawiają się *dla żaby*, a nie dla jej komponentów. Te uwagi stanowią dla McDowella podstawę do wyciągnięcia wniosku, iż to, co konstytuuje intencjonalne zjawiska z poziomu osobowego, samo powinno być poszukiwane na poziomie osobowym.

Aby przenieść rozważania McDowella na nieco inny grunt, spójrzmy na kolejną ilustrację takiego ujęcia dystynkcji osobowe–subosobowe. W kognitywistycznych rozważaniach nad naturą percepcji wzrokowej istnieją dwie wielkie tradycje, wyrastające, jak się na ogół sądzi, z dwóch fundamentalnie różnych i „konkurencyjnych" sposobów rozumienia natury widzenia. Pierwsza z nich ma swoje korzenie w pracy *Vision* Davida Marra (2010). Badacz ten próbował ująć percepcję wzrokową jako złożony, wieloetapowy proces obliczeniowy prowadzący do wytworzenia – na podstawie dwuwymiarowego obrazu rzucanego na siatkówkę oka – wewnętrznej, trójwymiarowej reprezentacji sceny wzrokowej. Druga tradycja wiąże się z postacią Jamesa Gibsona i jego ekologicznym podejściem do percepcji, które zostało wyrażone między innymi w pracy *The Ecological Approach to Visual Perception* (1979). W tradycji Gibsonowskiej traktuje się percepcję jako bezpośrednią, a nie mediowaną wewnętrznymi reprezentacjami czy (operującymi na reprezentacjach) procesami obliczeniowymi. W tym ujęciu percepcja polega nie tyle na przetwarzaniu informacji przez wewnętrzne mechanizmy obliczeniowe, co raczej na umiejętnym korzystaniu przez organizm z informacji bezpośrednio dostępnej w środowisku zewnętrznym (zawartej w świetle padającym na siatkówkę).

Z perspektywy tradycji Marrowskiej teoria Gibsona musi wydawać się fundamentalną pomyłką, która traktuje proces percepcji jako „magiczny" i niewymagający złożonej, wewnętrznej maszynerii. Co więcej, z tego punktu widzenia podejście drugiego autora wydaje się bezceremonialnie pomijać wiedzę o rzeczywistej złożoności neuronalnych mechanizmów percepcyjnych. Kiedy dysponujemy jednak odpowiednio przeprowadzoną dystynkcją osobowe–subosobowe, możemy zobaczyć, że tego rodzaju nieprzyjazna tradycji Gib-

sonowskiej konkluzja opiera się na konfuzji. Otóż można przyjąć, że projekty Gibsona i Marra są sformułowane na różnych poziomach organizacji i wyjaśniania systemu poznawczego (Bechtel, Abrahamsen 1993; McDowell 1994; Herschbach 2008). Marr zmierza do opisania mechanizmów wyjaśniających, jak wewnętrzna, subosobowa struktura systemów poznawczych umożliwia im percepcję wzrokową. Gibson ignoruje rolę tych mechanizmów nie dlatego, że odmawia im istnienia czy pełnienia ważnej roli, lecz dlatego, że jego projekt jest skoncentrowany na zupełnie innym poziomie organizacji systemu poznawczego, a przez to – na innym poziomie jego wyjaśniania. Jak to ujmuje McDowell, „krytycy Gibsona, niedysponujący odróżnieniem poziomów, czytają Gibsonowskie opisy systemów sensorycznych jak gdyby miały one spełniać tę samą funkcję intelektualną, którą pełnią ich własne teorie" (1994: 202). Tymczasem Gibson zmierza do dostarczenia szczegółowego wyjaśnienia, *na czym polega* percepcja wzrokowa – mówiąc językiem McDowella, co ją konstytuuje – z perspektywy percypującego organizmu (systemu poznawczego) jako całości. Opisuje on, co robi podmiot percepcji, a nie jak to robi. Gibson wyróżnia elementy środowiska, na które organizm jest percepcyjnie czuły, oraz wskazuje, jakie znaczenie mają dla tego organizmu dostępne w środowisku informacje[31]. Teoria tego autora nie mówi nic o wewnętrznych mechanizmach percepcji właśnie dlatego, że została sformułowana w całości na poziomie *osobowym*. Kiedy weźmiemy ten fakt pod uwagę, to widzimy – według McDowella – że ekologiczna teoria percepcji nie jest absurdalna, zważając na jej funkcję poznawczą. Widzimy też, że tradycje Marra i Gibsona wcale nie muszą być postrzegane jako wykluczające się wzajemnie[32]. Badacze ci są zaangażowani w zasadniczo różne projekty.

---

[31] Według Gibsona organizmy postrzegają przede wszystkim tak zwane afordancje, czyli potencjalne działania, jakie mogą być wykonane względem lub za pomocą percypowanych obiektów.

[32] Przykład ekologicznej teorii percepcji Gibsona pokazuje zarazem, że przyjęty tu sposób myślenia o naturze poziomu osobowego wcale nie implikuje, iż wyjaśnienia z tego poziomu – sformułowane za pomocą kategorii psychologii potocznej – są całkowicie niezależne od rozwoju wiedzy naukowej i naukowo

Żaba albo omówiony wcześniej wirtualny organizm są podmiotami prostych stanów percepcyjnych i mają niezbyt wyrafinowane preferencje, na podstawie których podejmują proste decyzje praktyczne. Teoria Gibsona koncentruje się na takim rodzaju percepcji wzrokowej, który nie jest specyficznie ludzki. Czy cały dotychczasowy wywód mówi zatem cokolwiek o naturze dystynktywnie ludzkich postaw propozycjonalnych – przekonań, pragnień, nadziei, oczekiwań, intencji i tak dalej – oraz o wyjaśnianiu za ich pomocą specyficznie ludzkich działań? Sądzę, że wbrew pozorom odpowiedź na to pytanie jest twierdząca. Psychologia potoczna, którą my, ludzie, aplikujemy do opisywania, wyjaśniania i przewidywania własnych oraz cudzych działań, jest skoncentrowana na poziomie osobowym, czyli systemowym. Pozwala ona na usytuowanie istot ludzkich w ramach zamieszkiwanego przez nie świata – na zasadzie podobnej, jak

niefalsyfikowalne. Twierdzę jedynie, że wyjaśnienia te mogą być rewidowane lub falsyfikowane tylko przez teorie naukowe sformułowane na tym samym, osobowym poziomie organizacji i wyjaśniania systemu poznawczego. Por. teza McCauleya (1986), że do eliminacji w nauce dochodzi tylko w wyniku zmian teoretycznych w obrębie jednego poziomu analizy, a nie w wyniku powstania nowych teorii sformułowanych na niższym poziomie. Można spekulować, że zdroworozsądkowa „teoria" percepcji wzrokowej (sformułowana na poziomie osobowym) głosi, iż percypowanie jest procesem pasywnym i polega na dysponowaniu szczegółowym, mentalnym obrazem percypowanej sceny (trzeba jednak zaznaczyć, że wyniki badań nad potocznymi przekonaniami dotyczącymi percepcji nie potwierdzają jednoznacznie takiej rekonstrukcji „zdroworozsądkowej" koncepcji widzenia; por. Schwitzgebel 2007). Niektóre nowsze, enaktywne koncepcje percepcji (por.: Noë 2002; O'Regan, Noë 2008) zmierzają do pokazania, że doświadczenie percepcyjne jest bardzo selektywne i ubogie treściowo, a proces percepcji stanowi – tak jak postulował Gibson – formę aktywnego, eksploracyjnego działania. Zakładam, że te nowe koncepcje, podobnie jak propozycja Gibsona, dotyczą poziomu osobowego. Z punktu widzenia takich teorii twierdzenie o istnieniu bogatych, oglądanych pasywnie obrazów mentalnych jest fałszywe. Jeśli (1) prawdziwa będzie rekonstrukcja potocznej teorii percepcji jako głoszącej istnienie szczegółowych obrazów percepcyjnych oraz (2) prawdziwe są tezy zwolenników ekologicznego/enaktywnego podejścia do percepcji, to mamy do czynienia z sytuacją, w której przekonanie wywiedzione z psychologii potocznej podlega falsyfikacji i rewizji w świetle wiedzy naukowej. Należy jednak podkreślić, że taka zależność jest możliwa tylko wtedy, gdy wiedza naukowa dotyczy poziomu osobowego, a nie mechanizmów subosobowych, na których temat, jak się wydaje, psychologia potoczna milczy.

w przypadku wirtualnego organizmu Beera czy żaby. Różnica polega na tym, że ludzki świat oraz ludzkie sposoby jego „zamieszkiwania" są niepomiernie bardziej złożone, niż w przypadku tamtych prostych systemów poznawczych. Można powiedzieć, że życie wszelkich ssaków, a szczególnie wysoko uspołecznionych naczelnych, jest dużo bardziej złożone, różnorodne i plastyczne. W przypadku ludzi mówimy jednak o istotach posługujących się językiem symbolicznym i mających za sobą tysiące lat ewolucji kulturowej. Bogactwo i złożoność osobowych kategorii intencjonalnych, za pomocą których opisujemy i wyjaśniamy ludzkie działania, jest odzwierciedleniem złożoności specyficznie ludzkich sposobów życia (por. Wittgenstein 2000). Do tej kwestii – a dokładniej do zagadnienia natury postaw propozycjonalnych w świetle bronionego tu ujęcia psychologii potocznej – powrócę w podrozdziale 5.3.

5.2.3. Konsekwencje mechanistycznego odczytania dystynkcji osobowe–subosobowe

Jak sądzę, broniona tu propozycja generuje szereg filozoficznie ważnych konsekwencji. W szczególności zaś każe ona zerwać z niektórymi zakorzenionymi we współczesnej filozofii umysłu i kognitywistyki założeniami dotyczącymi relacji międzypoziomowych, zobowiązań architekturalnych psychologii potocznej oraz przede wszystkim zależności między projektem naturalizacji intencjonalności a naturą subosobowych mechanizmów poznania. Można wymienić trzy takie „nieklasyczne" konsekwencje mechanistycznego odczytania opozycji osobowe–subosobowe. Po pierwsze, przyjmowane tu ujęcie wspomnianej dystynkcji pozwala na odrzucenie założenia o korespondencji osobowe–subosobowe, jednak bez zaprzeczania przy tym tezie o istnieniu postaw propozycjonalnych i ważnej eksplanacyjnej roli psychologii potocznej. Po drugie, broniona tu koncepcja pozwala na odrzucenie ZKOS w sposób, który nie pociąga za sobą antynaturalistycznych implikacji. Po trzecie, propozycja ta pozwala na sformułowanie wniosku, że istnieje zasadniczy rozdźwięk między naturą i funkcjami eksplanacyjnymi osobowych stanów intencjonalnych a naturą i funkcjami eksplanacyjnymi re-

prezentacji subosobowych. Skupmy się teraz na każdej z tych konsekwencji z osobna.

a) Odrzucenie ZKOS i mechanistyczna neutralność psychologii potocznej

Na tym etapie rozważań powinno być całkowicie jasne, że broniona tu przeze mnie propozycja wiąże się z kategorycznym odrzuceniem założenia korespondencji. Odrzucam ZKOS w sformułowaniu *epistemicznym*, ponieważ przyjmuję, że wyjaśnianie na poziomie osobowym spełnia zasadniczo inne funkcje i ma inną naturę niż wyjaśnianie na poziomie subosobowym; różnice te zostały wymienione w tabeli 3. Odrzucam jednak także *metafizyczne* sformułowanie ZKOS. Z punktu widzenia bronionej tu propozycji nie ma żadnych dobrych powodów, by uznawać, jakoby przekonania i pragnienia stanowiły wewnętrzne komponenty czy stany komponentów systemu poznawczego. Z mechanistycznej perspektywy nie musi zachodzić tego rodzaju korespondencja między poziomem osobowym a subosobowym. Właśnie z tego powodu można też powiedzieć, że psychologia potoczna jest *mechanistycznie neutralna*. Odwołuje się ona do własności wyższego rzędu, czyli własności systemowych, które nie muszą mieć żadnych odpowiedników na niższych poziomach organizacji systemu poznawczego. Ideę tę Susan Hurley wyraża w następujących słowach:

> To, czy struktury subosobowe korespondują ze strukturami, pojęciowymi lub innymi, z poziomu osobowego, stanowi każdorazowo kwestię empiryczną. Izomorfizm międzypoziomowy nie powinien być ani wymagany, ani negowany *a priori*. Treść z poziomu osobowego może pozostać dystynktywnie pojęciowa i normatywna, pomimo tego że wyjaśniamy umysły jako wyłaniające się na podstawie interakcji ucieleśnionych mózgów z ich środowiskiem, w tym środowiskiem społecznym. […] zachowanie systemu może być determinowane nieliniowymi relacjami zachodzącymi między czynnikami na niższym poziomie, chociaż jego struktura nie musi korespondować ze strukturą niższego poziomu (Hurley 2008: 21)[33].

[33] „Izomorfizm międzypoziomowy", o którym mówi Hurley w przytaczanym cytacie, to inaczej po prostu założenie korespondencji w wersji metafizycznej.

Skoro wyjaśnienia formułowane za pomocą psychologii potocznej odwołują się do własności systemowych, to okazuje się, że wcale nie muszą być one „rehabilitowane" przez to, co Godfrey-Smith (2004) nazwał faktami architekturalnymi. Poprawność wyjaśniania i opisywania ludzi jako racjonalnych posiadaczy przekonań, pragnień i innych postaw propozycjonalnych jest w pewnym stopniu niezależna od faktów zachodzących na poziomie subosobowym. Bycie podmiotem przekonania nie wymaga posiadania „w głowie" jakiegoś stanu subosobowego, który moglibyśmy z tym przekonaniem zidentyfikować; to samo dotyczy wszystkich innych postaw propozycjonalnych. Do wątku tego powrócę jeszcze w punkcie (c).

Jeśli przypomnimy sobie przedstawioną w rozdziale 2 (podrozdział 2.2) dyskusję wielopoziomowości mechanizmów, to szybko okaże się, że powyższe stwierdzenia nie zawierają nic, co mogłoby zostać uznane za zaskakujące czy problematyczne w świetle wiedzy na temat natury mechanizmów i wyjaśnień mechanistycznych. Przypisywanie własności *systemowych* zlokalizowanym *komponentom* systemu czy mechanizmu stanowi na ogół błąd lub co najwyższej narzędzie heurystyczne użyteczne na wstępnym etapie badań. Własności systemowe nie są na ogół egzemplifikowane przez wewnętrzne komponenty systemów czy mechanizmów (Bechtel, Abrahamsen 1993; Bechtel, Richardson 1993; Wimsatt 2006a; por. także: Bennett, Hacker 2003; Poczobut 2009). Nieagregatywne systemy jako całości mogą egzemplifikować własności, które nie przysługują ich częściom składowym. Z tego też powodu kategorie (predykaty), za pomocą których jest opisywany system i jego interakcje ze środowiskiem, różnią się od kategorii, za pomocą których opisuje się jego wewnętrzną, mechanistyczną strukturę. Co więcej, przypisywanie różnego rodzaju własności systemowych (1) jest możliwe pod nieobecność wiedzy na temat mechanizmów stojących u podstaw tych własności; (2) nie przesądza o tym, jakie to mechanizmy (Bechtel, Abrahamsen 1993; Bechtel 2008: 146–147, 155–157). Z perspektywy mechanistycznej wymienione twierdzenia dotyczą w zasadzie wszystkich złożonych, nieagregatywnych systemów występujących w przyrodzie. Tutaj twierdzę po prostu, że systemy *poznawcze* nie są tu wyjątkiem. Odrzucając ZKOS, nie postuluję jakiejś „tajemni-

czej", radykalnej metafizyki postaw propozycjonalnych, lecz wpisuję się jedynie w mechanistyczny sposób myślenia o międzypoziomowych zależnościach w złożonych, nieagregatywnych systemach fizycznych.

Zanim będzie można przejść dalej – warto nadać bardziej konkretną postać twierdzeniu o mechanistycznej neutralności poziomu osobowego. Wyobraźmy sobie za Lynne Baker (2001), że Jan jest członkiem rady szkolnej w prywatnym liceum. Mamy wszelkie powody, by sądzić, iż jest on przeciwnikiem podniesienia szkolnego czesnego. Sprzeciwiał on się podniesieniu czesnego w trakcie spotkania rady. Napisał też list do dyrektorki, w którym wypowiedział się przeciwko wydatkom zbytnio obciążającym jego zdaniem budżet szkoły i przez to mogącym prowadzić do podwyższenia czesnego. Pewnego dnia odbyło się spotkanie rady szkolnej, na którym głosowano budżet na najbliższy rok. Jan w trakcie spotkania kilkakrotnie wyraził przekonanie, że przyjęcie tego budżetu doprowadzi do konieczności podniesienia czesnego. W końcu głosuje on przeciw przyjęciu tego budżetu. Biorąc pod uwagę przytoczone świadectwa behawioralne, wydaje się niekontrowersyjne stwierdzenie, że każde z wymienionych działań Jana jest wynikiem żywionego przez niego pragnienia, by czesne nie zostało podniesione.

Wyobraźmy sobie teraz za Baker (2001), iż dysponujemy kompletną mapą pozwalającą nam zobaczyć z dokładnością do jednego neuronu, co działo się wewnątrz mózgu Jana w trakcie wykonywania każdego z działań wymienionych powyżej. Zapytajmy za tą autorką: czy poprawność (prawdziwość) wyjaśnienia działań Jana za pomocą osobowego stanu intencjonalnego – pragnienia, aby czesne nie zostało podniesione – wymaga tego, by każde z tych działań było przyczynowo wywołane rodzajowo identycznym stanem czy strukturą subosobową?[34] Filozofowie przyjmujący założenie korespondencji odpowiedzieliby twierdząco. Zgodnie z ZKOS poprawność wy-

[34] Robert Piłat zagadnienie to formułuje następująco: „Powiedzmy, że po raz n-ty w swoim życiu stwierdzam »Myślę, że ptaki już odleciały na zimę«. Czy to znaczy, że jestem po raz n-ty w tym samym stanie umysłowym, a przynajmniej w stanie umysłowym zachowującym pewien identyczny rdzeń w różnych okolicznościach, w których powtarza się mój sąd? Od czego zależałaby

jaśnień działań Jana na poziomie osobowym zależy od tego, czy każde z tych działań może być wyjaśnione przez jakiś jeden (rodzajowo czy typicznie) stan neuronalny czy neuroobliczeniowy, który stanowi subosobowy odpowiednik pragnienia, aby czesne nie zostało podniesione. Oznacza to, że powinna istnieć zbieżna z psychologią potoczną, subosobowa taksonomia stanów mentalnych umożliwiająca ujęcie każdego z działań Jana jako skutku aktywności rodzajowo czy typicznie identycznego stanu wewnętrznego, na przykład zdania w języku myśli albo biologicznego stanu informacyjnego.

Z przyjmowanej tu mechanistycznej perspektywy istnienie tego rodzaju subosobowego odpowiednika pragnienia Jana jest co prawda empirycznie możliwe. Jednakże (1) wydaje się to mało prawdopodobne, biorąc pod uwagę kierunek rozwoju kognitywistyki na przestrzeni ostatnich dziesięcioleci; (2) nie jest to (empirycznie ani pojęciowo) *konieczne* do tego, aby wyjaśnienia działań Jana za pomocą jego pragnienia były prawomocne (poprawne, prawdziwe). Twierdzę, że nie musi istnieć – i prawdopodobnie nie istnieje – jeden (typicznie) stan czy jedna własność neurofizjologiczna, neurooblicczeniowa, a mówiąc szerzej *subosobowa*, która odpowiadałaby przyczynowo za każde z działań Jana. Psychologia potoczna jest mechanistycznie neutralna; posługiwanie się nią w prawomocny sposób nie wymaga zachodzenia takiej korespondencji międzypoziomowej.

Przyjmijmy zatem, iż fakt, że za wszystkie (wymienione wyżej) działania Jana odpowiada jedno pragnienie, jest widoczny *wyłącznie* z perspektywy, jaką daje nam psychologia potoczna. Okazuje się, że nie istnieje żaden inny niż osobowy punkt widzenia czy aparat pojęciowy, z którego perspektywy aktywność Jana stanowiłaby wynik rodzajowo identycznego stanu mentalnego. Dla zwolennika ZKOS taki scenariusz stanowi punkt wyjścia do wyciągnięcia konkluzji eliminatywistycznej. Jednakże z perspektywy bronionego tu stanowiska taki obrót spraw nie zagraża istnieniu i roli eksplanacyjnej pragnienia żywionego przez Jana. Pragnienie, aby czesne nie zostało podniesione, jest kwestią tego, jak Jan interreaguje ze swoim środo-

identyczność tego stanu umysłowego? Czy te domniemane stany umysłowe byłyby ściśle skorelowane ze stanami mózgu?" (1999: 81).

wiskiem społeczno-kulturowym, a nie – posiadania „w głowie" określonej struktury subosobowej. To znaczy, że żywienie pragnienia zależy od praktycznej, działaniowej zależności między Janem a pewnym makro-aspektem jego otoczenia, a nie od wewnętrznej budowy jego mózgu (por. Bechtel, Abrahamsen 1993; Clark 1993: 189–219; McDowell 1994; Dennett 2008). Wyjaśnienia psychologii potocznej koncentrują się na tego rodzaju zależnościach, występujących na poziomie systemu poznawczego jako całości zaangażowanej w interakcje ze swoim środowiskiem[35].

b) Trzecia droga: między ZKOS a antynaturalistycznym separacjonizmem

Kolejna ważna konsekwencja przyjętego tu stanowiska w sprawie dystynkcji osobowe–subosobowe dotyczy dylematu wspomniane-

[35] Aby dodatkowo rozjaśnić to twierdzenie, spróbuję wyrazić stojącą za nim ideę w jeszcze inny sposób. Sądzę, że postawy propozycjonalne są *własnościami dyspozycyjnymi* (wiązkami własności dyspozycyjnych) systemu poznawczego jako całości; działania wynikające z posiadania określonej postawy stanowią więc manifestacje pewnej własności dyspozycyjnej (dyspozycji do działania, myślenia i odczuwania w sposób charakterystyczny dla podmiotów tej postawy; por. podrozdział 5.3). Otóż odrzucam tezę, jakoby różne działania (na przykład zabranie ze sobą parasola albo wypowiedzenie zdania „Pada deszcz") stanowiły manifestację rodzajowo identycznego stanu intencjonalnego o treści *p* (na przykład przekonania, że pada deszcz) dlatego, że każde z nich jest przyczynowo wywołane rodzajowo identycznym stanem subosobowym o treści *p* (na przykład jakimś neuronalnym czy neuroobliczeniowym stanem niosącym treść „Pada deszcz"). Z bronionej tu perspektywy jest dokładnie na odwrót. Poszczególne stany subosobowe mogą być uznane za przyczynową podstawę działań stanowiących manifestację rodzajowo identycznego przekonania tylko wtedy, gdy działania przez nie wywołane stanowią manifestacje jednej i tej samej własności dyspozycyjnej. Tych subosobowych stanów może nie łączyć nic więcej, żadna fizjologiczna czy obliczeniowa własność, którą wszystkie by posiadały. Działania, w jakich manifestuje się przekonanie, że *p*, mogą mieć u swoich podstaw wysoce heterogeniczną klasę stanów subosobowych (inny rodzajowo stan subosobowy wywołuje zabranie parasola, inny odpowiada za wypowiedzenie zdania „Pada deszcz" i tak dalej). Te subosobowe stany mogłyby zostać zaliczone do jednego rodzaju czy typu, tylko gdybyśmy *presuponowali* opis w kategoriach psychologii potocznej. Postawy propozycjonalne przejawiają się we wzorcach działania widocznych jedynie z perspektywy psychologii potocznej (por.: Clark 1993: 189–219; Dennett 2008).

go w sekcji 5.1.3. Chodzi o rzekomą alternatywę: albo akceptujemy założenie korespondencji osobowe–subosobowe, odrzucamy autonomię poziomu osobowego i zachowujemy dzięki temu naturalizm, albo odchodzimy od naturalizmu i negujemy związki eksplanacyjne między poziomem osobowym i subosobowym. Wydaje się, że proponowana tu mechanistyczna interpretacja owej opozycji stanowi kompromis między tymi dwoma skrajnymi biegunami. Na czym on polega? Otóż ta propozycja pozwala oddać sprawiedliwość poziomowi osobowemu jako swoistemu i (częściowo) autonomicznemu, jednak w sposób, który nie prowadzi do antynaturalistycznego separacjonizmu.

Akceptuję tu szereg pozornie sepracjonistycznych tez (por.: Davidson 1992; Hornsby 2000) głoszących, że: (1) między naturą i funkcjami epistemicznymi wyjaśnień na poziomie osobowym i subosobowym zachodzą ważne różnice; (2) można formułować poprawne i kompletne wyjaśnienia osobowe, które w ogóle nie powołują się na fakty subosobowe; (3) między psychologią potoczną a kognitywistyką nie może zajść interteoretyczna redukcja, rozumiana „klasycznie", w sposób inspirowany modelem N-D wyjaśniania; (4) orzekanie osobowych predykatów mentalnych o stanach czy strukturach subosobowych stanowi błąd kategorialny czy też błąd mereologiczny (por. Bennett, Hacker 2003). Sądzę jednak, że akceptacja tych tez wcale nie implikuje antynaturalizmu głoszącego *całkowitą* autonomię poziomu osobowego. Odróżnijmy bowiem dwa twierdzenia:

> **(T1)** Poziom osobowy i subosobowy wyznaczają dwa różne, wzajemnie autonomiczne sposoby opisywania oraz wyjaśniania jednego i tego samego bytu (umysłu czy systemu poznawczego).
> **(T2)** Poziom osobowy i subosobowy wyznaczają dwa różne, wzajemnie autonomiczne sposoby opisywania oraz wyjaśniania działań jednego i tego samego bytu (umysłu czy systemu poznawczego) *na różnych poziomach jego mechanistycznej organizacji*.

Jak się wydaje, (T1) niesie antynaturalistyczne konsekwencje. Głosi ono, że dysponujemy dwoma całkowicie niezależnymi sposobami konceptualizowania i wyjaśniania jednego i tego samego bytu. Jednak bronione tu stanowisko jest wyrażone w (T2), a nie (T1). Mówiąc obrazowo, w świetle (T2) wyjaśnienia z poziomu osobowego i subosobowego nie są konkurencyjnymi „lokatorami" tego samego piętra organizacji systemu poznawczego, lecz „mieszkańcami" różnych pięter. Częściowa autonomia poziomu osobowego względem subosobowego będzie całkowicie naturalna, jeśli wziąć pod uwagę fakt, że oba poziomy wyjaśniania koncentrują się na różnych piętrach mechanistycznej organizacji systemu poznawczego.

W rozdziale 2 (sekcja 2.2.2) zostało powiedziane, że dociekania skoncentrowane na wyższym poziomie organizacji w zasadzie zawsze cechują się pewnym stopniem autonomii względem dociekań dotyczących mechanizmów niższego poziomu. Wyższy poziom jest opisywany na ogół za pomocą innego języka i są formułowane na jego temat generalizacje, które nie zostają odzwierciedlone na niższych poziomach. Fizjolog może badać i opisywać funkcjonalne zależności między narządami wewnętrznymi nowoodkrytego gatunku organizmów – w tym także formułować (za pomocą swoistych dla tego poziomu organizacji kategorii) naukowo wartościowe generalizacje na temat tych zależności – pod nieobecność wiedzy na temat mechanizmów odpowiadających za funkcjonowanie i interakcje owych narządów. Fizjolog może też w takiej sytuacji generować horyzontalne wyjaśnienia przyczynowe, pokazujące na przykład, że narząd *y* znajduje się w momencie *t* w określonym stanie dlatego, że w momencie *t-1* w określonym stanie znalazł się narząd *x*, z którym *y* jest w określony sposób sprzężony przyczynowo i funkcjonalnie (por. Bechtel, Abrahamsen 1993). Koncentrując się na konkretnym, narządowym poziomie organizacji, dociekania fizjologa są w pewnym zakresie autonomiczne względem dociekań dotyczących mechanizmów niższego poziomu.

Twierdzę, że autonomia poziomu osobowego jest podobnego rodzaju. Wynika ona nie z faktu istnienia radykalnie niewspółmiernych sposobów wyjaśniania jednego obiektu badań, ale z tego, że wyjaśnienia osobowe dotyczą innego poziomu organizacji niż suboso-

bowe[36]. Z perspektywy mechanicyzmu to całkowicie naturalny stan rzeczy. Jeśli mechanistyczne ujęcie dystynkcji osobowe–subosobowe jest zasadne, to *powinniśmy* oczekiwać, że poziom osobowy będzie posługiwał się specyficznymi kategoriami i generalizacjami oraz że nie będzie zachodziła międzypoziomowa korespondencja. *Powinniśmy* też oczekiwać, że jest możliwe dysponowanie kompletnymi, horyzontalnymi wyjaśnieniami osobowymi, które nie odwołują się („wertykalnie") do struktur z poziomu subosobowego. Wreszcie *powinniśmy* porzucić oczekiwanie, że wyższy poziom będzie podlegał interteoretycznej redukcji do poziomu niższego; mechanistyczna redukcja zjawisk z wyższego poziomu jest zawsze częściowa i zasadniczo różni się od redukcji rozumianej jako dedukcja praw teorii redukowanej z praw teorii redukującej. Każde z tych oczekiwań wpisuje się w mechanistyczne spojrzenie na naturę poziomów wyjaśniania oraz relacji między nimi. Jak sądzę, autonomia poziomu osobowego nie jest kłopotliwa dla naturalistów.

Przyjmowane tu stanowisko nie wyłamuje się z naturalizmu także ze względu na fakt, że na jego gruncie autonomia poziomu osobowego jest jedynie częściowa. Nie neguję tu istnienia ważnych związków eksplanacyjnych między poziomem osobowym a subosobowym. Więcej – sądzę, że zasadniczą funkcją kognitywistyki jest odkrywanie i opisywanie tego rodzaju związków. Kognitywistyka w znacznym stopniu zajmuje się formułowaniem wertykalnych wy-

[36] Choć teza ta wymagałaby rozwinięcia i szczegółowej argumentacji, sądzę, że przy przyjęciu bronionej tu perspektywy nawet normatywność zjawisk umysłowych na poziomie osobowym nie stanowi dla naturalisty wyzwania, któremu nie mógłby on podołać. Racjonalne działania i procesy inferencyjne sprawiają dla naturalisty problem wtedy, gdy próbuje on „szukać" tych zjawisk na poziomie subosobowej architektury umysłu. Racjonalność jest tymczasem własnością „zlokalizowaną" na najwyższym poziomie organizacji systemu poznawczego – poziomie osobowym. Podejrzewam, że zrozumienie natury racjonalności będzie wymagało nie tyle spojrzenia w dół, na wewnętrzną organizację systemu poznawczego, co raczej spojrzenia w górę, na relacje tego systemu ze środowiskiem społeczno-kulturowym, a także na naturę praktyk społecznych, w tym tych specyficznie ludzkich (por.: Toribio 1998; Hurley 2008; Gładziejewski 2012). Szukanie racjonalności na poziomie subosobowym jest być może skazane na niepowodzenie.

jaśnień zjawisk *osobowych*, które to wyjaśnienia odwołują się do *subosobowych* mechanizmów odpowiadających za te zjawiska. Mówiąc wprost, celem nauk kognitywnych pozostaje wyjaśnianie osobowych eksplanandów za pomocą subosobowych eksplanansów. Eksplanandami tymi nie są postawy propozycjonalne. Te ostatnie nie są jednak jedynymi zjawiskami występującymi na poziomie osobowym. Przedmiotem kognitywistycznych wyjaśnień są własności systemu poznawczego jako całości, jednak nie przekonania i pragnienia. Jakie to zatem własności? Jak zostało wcześniej wspomniane (por. rozdział 2, szczególnie sekcje 2.1.1 oraz 2.3.1), zasadniczy przedmiot mechanistycznych wyjaśnień kognitywistyki stanowią zdolności czy kompetencje systemu poznawczego. Zauważmy, że tego rodzaju zdolności przypisuje się systemowi jako całości. Są one własnościami tego systemu, a nie jego części. To nie kora wyspy posiada zdolność do empatyzowania, lecz podmiot (system poznawczy); to nie kora wzrokowa, lecz podmiot (system poznawczy) percypuje świat (por. Bennett, Hacker 2003). Pamiętanie relacji przestrzennych, podejmowanie decyzji, rozumienie mowy, formowanie pojęć, kategoryzacja percepcyjna i inne zdolności stanowiące eksplananda kognitywistyki są „zlokalizowane" na poziomie systemowym (osobowym), a wyjaśniają je subosobowe mechanizmy poznawcze (por.: Bechtel, Abrahamsen 1993; Herschbach 2008). Na poziomie osobowym określamy, co „potrafi" system poznawczy, jakimi zdolnościami dysponuje. Kognitywistyka pokazuje, jak działanie wewnętrznych, subosobowych mechanizmów umożliwia systemowi poznawczemu posiadanie tych zdolności. Poziom osobowy dostarcza zatem kognitywistyce eksplanandów, które ta wyjaśnia na poziomie subosobowym[37]. Tym samym między poziomem osobowym a subosobowym nie występuje żadna eksplanacyjna „przepaść".

[37] Zauważmy na marginesie, że takie ujęcie nie tylko pokazuje eksplanacyjną wagę tego, co subosobowe, dla tego, co osobowe, ale też niezbywalność samych zjawisk osobowych. Naukowcy potrzebują poziomu osobowego, aby móc wyróżnić i scharakteryzować wyjaśniane przez siebie eksplananda. Bez poziomu osobowego nie wiedzielibyśmy, co powinno być wyjaśnione (de Pinedo, Noble 2008).

c) Naturalizowanie intencjonalności a reprezentacje osobowe i subosobowe

Zwróćmy teraz uwagę na kwestię relacji między problemem naturalizacji intencjonalności a problemem eksplanacyjnego statusu reprezentacji w kognitywistyce. Założenie korespondencji każe postrzegać te dwa zagadnienia jako ściśle powiązane. Z punktu widzenia ZKOS – jeśli chcemy znaleźć miejsce dla postaw propozycjonalnych w naukowym obrazie świata, musimy pokazać, (1) że postawy te mogą być uznane za identyczne z określonymi subosobowymi stanami organizmów; (2) że tego rodzaju subosobowe odpowiedniki postaw propozycjonalnych rzeczywiście widnieją w poprawnych kognitywistycznych wyjaśnieniach zjawisk poznawczych. Zwolennicy projektu naturalizacji intencjonalności na ogół starają się pokazać, że (1) jest prawdą. Z kolei eliminatywiści – a w każdym razie eliminatywiści architekturalni – zmierzają do pokazania, że (2) jest fałszem. Obie strony tego sporu uzależniają jednak istnienie postaw propozycjonalnych od tego, czy w kognitywistyce zwycięży ściśle określona forma reprezentacjonizmu – mianowicie głosząca, że działanie systemu poznawczego opiera się na aktywności struktur subosobowych zachowujących intencjonalne/funkcjonalne własności postaw propozycjonalnych.

Odrzucając ZKOS, odrzucam zarazem przekonanie, że postawy propozycjonalne muszą albo okazać się identyczne ze stanami czy strukturami subosobowymi, albo zniknąć z naukowego obrazu świata. Z bronionej tu perspektywy takie przekonanie stanowi wynik „pomieszania" poziomów organizacji i wyjaśniania, które powinny być od siebie wyraźnie i jednoznacznie odróżniane. Mechanistyczna neutralność psychologii potocznej sprawia, iż kwestię istnienia postaw propozycjonalnych w świecie fizycznym oraz kwestię użyteczności pojęcia reprezentacji dla kognitywistyki należy uważać za zagadnienia od siebie niezależne. Rola eksplanacyjna postaw propozycjonalnych fundamentalnie różni się od roli eksplanacyjnej reprezentacji widniejących w mechanistycznych, subosobowych wyjaśnieniach kognitywistyki. Nic zatem zaskakującego, jeśli okaże się, że postawy propozycjonalne nie mają subosobowych odpowiedników i przez to nie mogą odgrywać roli w mechanistycznych wyja-

śnieniach zjawisk poznawczych. Nie na tym polega ich funkcja eksplanacyjna. Szukanie przekonań i pragnień w mózgu lub negowanie ich istnienia na podstawie faktów dotyczących subosobowej architektury systemu poznawczego stanowi za każdym razem wynik tego samego błędu – pomylenia dwóch zasadniczo różnych narzędzi eksplanacyjnych, specyficznych dla różnych poziomów wyjaśniania.

Warto doprecyzować twierdzenie o dwóch „zasadniczo różnych narzędziach eksplanacyjnych". Otóż jeśli przedstawiony tu wywód jest poprawny, to istnieją dwa bardzo odmienne rodzaje reprezentacji mentalnych. Pierwszy z nich to osobowe stany intencjonalne; nazwijmy je „reprezentacjami osobowymi". Są one (1) własnościami wyższego rzędu (systemowymi), przysługującymi systemom poznawczym jako całościom; (2) pełnią funkcję w horyzontalnych wyjaśnieniach poszczególnych działań systemów poznawczych. Drugi rodzaj reprezentacji to określone wewnętrzne struktury systemu poznawczego. Nazwijmy takie reprezentacje „subosobowymi". Są one (1) funkcjonalnie scharakteryzowanymi komponentami mechanizmów poznawczych; (2) pełnią funkcję w wertykalnych, mechanistycznych wyjaśnieniach zdolności poznawczych. W świetle tez bronionych w tej pracy reprezentacjami subosobowymi są modele mentalne. To właśnie ich – w przeciwieństwie do przekonań i pragnień – możemy szukać w mózgu.

Biorąc pod uwagę różnice zachodzące między reprezentacjami osobowymi a subosobowymi, możemy dojść do dwóch wniosków. Po pierwsze – co zostało już powiedziane – istnienie i wartość eksplanacyjna reprezentacji osobowych nie zależy od istnienia i wartości eksplanacyjnej tych subosobowych. Twierdząc przeciwnie, popełniamy błąd charakterystyczny dla zwolenników tradycyjnie pojmowanej naturalizacji intencjonalności i eliminatywistów. Po drugie, fakt, iż postawy propozycjonalne występują tylko na poziomie osobowym, nie czyni pojęcia reprezentacji subosobowych w jakiś sposób filozoficznie problematycznym. Przekonania i pragnienia występują jedynie na poziomie osobowym, co nie oznacza jednak – wbrew temu, co twierdzą chociażby McDowell (1994) czy Jennifer Hornsby (2000) – że każda sytuacja, w której ktoś przypisuje treść intencjonalną stanowi subosobowemu, jest wyrazem filozoficznej konfuzji.

Twierdzenie o istnieniu reprezentacji na poziomie subosobowym będzie jak najbardziej sensowne i niemetaforyczne – pod warunkiem, że zdajemy sobie sprawę, iż chodzi o reprezentacje zasadniczo odmienne (zarówno metafizycznie, jak i pod względem pełnionych ról eksplanacyjnych) od tych, które znajdujemy na poziomie osobowym.

Powtórzmy raz jeszcze najważniejszy wniosek płynący z tych rozważań: prawomocność praktyki opisywania oraz wyjaśniania działań za pomocą przekonań i pragnień (szerzej: reprezentacji osobowych) jest niezależna od tego, czy (oraz jakie) rodzaje reprezentacji subosobowych okażą się wartościowe ze względu na epistemiczne cele kognitywistów. Sukces lub porażka reprezentacjonizmu w kognitywistyce nie decyduje o metafizycznych losach postaw propozycjonalnych[38]. Osobowe fakty dotyczące przekonań i pragnień nie

[38] Nie znaczy to, że nie mamy tu potencjalnie do czynienia z ciekawym zagadnieniem empirycznym. Załóżmy, że modele wewnętrzne stanowią (jedyny) wartościowy eksplanacyjnie rodzaj reprezentacji subosobowych. Skoro są one reprezentacjami subosobowymi, to aby spełniać swoją rolę eksplanacyjną, nie muszą mieć żadnych własności, intencjonalnych czy funkcjonalnych, charakterystycznych dla przekonań czy pragnień. Ich rola eksplanacyjna jest zasadniczo różna od tej, którą pełnią przekonania i pragnienia. Niewykluczone jednak, iż badania empiryczne wykażą ciekawe zależności między osobowymi stanami intencjonalnymi a modelami mentalnymi. Piłat (1999: 87–88) formułuje hipotezę, zgodnie z którą u podstaw dyspozycji wynikających z posiadania postawy propozycjonalnej o określonej treści stoi zawsze jeden model mentalny. Jeśli hipoteza ta jest poprawna, to okazałoby się, że zachodzi korespondencja między poziomem osobowym (postawą propozycjonalną o określonej treści) a subosobowym (modelem mentalnym odpowiadającym tej postawie). (Zakładam rzecz jasna, że Piłat przystałby na tezę, iż modele mentalne są subosobowe). Model wewnętrzny stanowiłby stały subosobowy „rdzeń" dla szeregu dyspozycji związanych z posiadaniem określonego stanu intencjonalnego (na przykład przekonania, że ptaki już odleciały na zimę). Trzeba jednak podkreślić, że choć jestem otwarty na tego rodzaju propozycję jako *hipotezę empiryczną*, to stoję na stanowisku, iż posiadanie przekonania jest *konstytuowane* (w sensie McDowella) przez posiadanie określonej własności dyspozycyjnej (wiązki takich własności), wyróżnianej na poziomie osobowym. Przypisanie komuś przekonania nie niesie ze sobą zobowiązań architekturalnych (por. podrozdział 5.3). Ludzie byliby podmiotami przekonań, pragnień, oczekiwań i innych postaw propozycjonalnych, nawet gdyby hipoteza Piłata okazała się fałszywa. Być może posiadanie przekonania ma u swoich podstaw model mentalny, jednak wbrew temu

stoją przed trybunałem faktów dotyczących reprezentacji subosobowych. Nie znaczy to, że projekt naturalizacji intencjonalności okazuje się całkowicie pozbawiony sensu. Moim celem nie jest odmówienie filozoficznej doniosłości pytaniu o miejsce osobowych stanów mentalnych w świecie fizycznym. Podejmuję to pytanie w następnym podrozdziale. Tu chodzi mi jedynie o zasugerowanie, że nie odnajdziemy odpowiedzi, szukając naturalnych warunków wystarczających do tego, by stan subosobowy mógł zostać utożsamiony z osobowym stanem intencjonalnym.

5.3. Postawy propozycjonalne jako własności wyższego rzędu

5.3.1. Postawy propozycjonalne, zobowiązania architekturalne i „miękka" naturalizacja. Ustalenia wstępne

Wróćmy teraz do kwestii natury postaw propozycjonalnych i ich miejsca w świecie fizycznym. Rezygnując z założenia korespondencji, odrzuca się jedno potencjalne rozwiązanie tego problemu, lecz nie generuje się przy tym żadnej pozytywnej alternatywy. Spróbuję teraz naszkicować taką alternatywną teorię. Proponowana tu odpowiedź na pytanie o naturę przekonań i pragnień została już wstępnie wyrażona w sekcji 5.2.1. Zgodnie z nią postawy propozycjonalne są własnościami wyższego rzędu, przysługującymi systemom poznawczym jako całościom, a nie subosobowym komponentom tych systemów. Precyzyjniej, postawy propozycjonalne są własnościami *dyspozycyjnymi* systemów poznawczych. Celem prezentowanego podrozdziału będzie rozwinięcie tej ogólnej myśli. Zanim jednak przejdę do właściwych rozważań, należy poczynić kilka ważnych uwag wstępnych.

Po pierwsze, trzeba zaznaczyć, że przedstawione tu rozważania mają charakter szkicu. Moim celem nie jest zbudowanie teorii

autorowi (Piłat 1999: 88) nie sądzę, by *oznaczało* ono (było konstytuowane przez) posiadanie „w głowie" pewnego modelu mentalnego.

postaw propozycjonalnych, która pretendowałaby do miana kompletnej. Chodzi raczej o zarysowanie kierunku, w jakim należałoby zmierzać, jeśli chce się stworzyć koncepcję osobowych stanów intencjonalnych jako własności wyższego rzędu, to znaczy jako własności „zlokalizowanych" na poziomie systemu poznawczego jako całości. Chcę pokazać, że taka teoria jest osiągalna, a także wskazać, jaki kształt może ona przyjąć. Inaczej mówiąc, chcę tu przede wszystkim wykazać, że jest możliwe wypracowanie takiego ujęcia natury postaw propozycjonalnych, które wpisuje się w bronioną przeze mnie koncepcję relacji psychologii potocznej do kognitywistyki. Muszę też wyraźnie zaznaczyć, że nie będę tu rozwijał koncepcji oryginalnej. Chcę raczej skorzystać z istniejących już w literaturze filozoficznej zasobów teoretycznych. Wykorzystam przede wszystkim dyspozycyjno-fenomenalną teorię postaw propozycjonalnych Erica Schwitzgebela (2002, 2013), choć nawiążę także do teoretyczno-pomiarowej teorii postaw rozwijanej przez Roberta Matthewsa (2007, 2011). W celu pokazania, że postawy propozycjonalne rozumiane jako własności dyspozycyjne mogą być efektywne przyczynowo, skorzystam z interwencjonistycznej teorii przyczynowości Jamesa Woodwarda (2003).

Po drugie, należy bardzo precyzyjnie określić, jaką funkcję teoretyczną ma spełniać koncepcja postaw propozycjonalnych, o którą mi tu chodzi. Przypomnijmy sobie McDowella rozróżnienie na (1) wyjaśnienia wskazujące subosobowe (mechanistyczne) warunki umożliwiające fenomeny z poziomu osobowego oraz (2) konstytutywne wyjaśnienia tych fenomenów. O ile te pierwsze eksplanacje dotyczą tego, jak (dzięki jakim mechanizmom) są możliwe zjawiska osobowe, o tyle te drugie zmierzają do określenia, czym są te zjawiska (co je konstytuuje). Jak mógłby to ująć McDowell, wyjaśnienia konstytutywne pokazują, jakim rodzajem „kompetentnego zamieszkiwania świata" są określone fenomeny (zdolności, funkcje, stany, własności) z poziomu osobowego. Otóż przedstawiona tu propozycja ma taki charakter: jest ona konstytutywnym wyjaśnieniem (szkicem konstytutywnego wyjaśnienia) postaw propozycjonalnych. Będę zmierzać do pokazania, czym jest albo na czym polega posiadanie przekonania czy pragnienia. Zgodnie z omawianą tu pozycją

postawy propozycjonalne są własnościami wyższego rzędu systemów poznawczych. To system poznawczy jako całość, a nie żaden jego komponent, jest posiadaczem przekonań i pragnień. Odrzucam tu ideę, jakoby atrybucja przekonań i pragnień niosła ze sobą zobowiązania architekturalne, to znaczy wymagała, by w ramach systemu istniały jakieś komponenty czy stany komponentów, które moglibyśmy utożsamiać z postawami propozycjonalnymi podmiotu. Sądzę, że konstytutywne wyjaśnienie postaw propozycjonalnych nie będzie wymagało „spoglądania" w dół, na poziom subosobowy. Dlatego poszukuję teorii, która będzie pozwalała ująć naturę postaw propozycjonalnych w sposób abstrahujący od tego, czy da się odnaleźć jakieś ich subosobowe odpowiedniki.

Posługując się dystynkcją zaproponowaną przez Schwitzgebela (2013), można też powiedzieć, że zgodnie z omawianą tu koncepcją postawy propozycjonalne są powierzchownymi (*superficial*), a nie głębokimi (*deep*) własnościami systemu poznawczego. System S posiada głęboką właśność nie tylko ze względu na powierzchowne (zewnętrzne) wzorce (*patterns*), ale także ze względu na obecność określonych cech wewnętrznych. Własność powierzchowną system S posiada jedynie na podstawie zewnętrznych wzorców. Kwalifikacja własności jako głębokiej lub powierzchownej będzie zawsze zrelatywizowana do tego, co wyróżniamy jako klasę „powierzchownych" czy zewnętrznych wzorców (Schwitzgebel 2013). Przyjmijmy zatem, że w przypadku systemów poznawczych powierzchowne wzorce zostają wyróżnione na poziomie osobowym, czyli na poziomie interakcji systemu poznawczego jako całości z zamieszkiwanym przez niego środowiskiem. Przyjmijmy też, że powierzchowne konstytutywne wyjaśnienie zjawiska z poziomu osobowego to konstytutywne wyjaśnienie tego zjawiska za pomocą pewnej własności powierzchownej. Mając na uwadze wszystkie te rozróżnienia, można już precyzyjnie wyrazić cel prowadzonych tu rozważań. Chcę zarysować teorię dostarczającą *powierzchowne konstytutywne wyjaśnienie postaw propozycjonalnych*[39].

[39] Nie twierdzę, że głębokie konstytutywne wyjaśnienia zjawisk z poziomu osobowego są niemożliwe, albo że wyjaśnienia powierzchowne zawsze będą tymi

Po trzecie, broniąc tego rodzaju powierzchownej koncepcji postaw propozycjonalnych, moim celem nie jest negowanie twierdzenia, że byciem podmiotem przekonań i pragnień wymaga spełnienia pewnych minimalnych, stosunkowo trywialnych warunków architekturalnych (dotyczących poziomu subosobowego). Jakie to warunki? Jak się wydaje, można wymienić dwa. Pierwszy głosi po prostu, że system poznawczy będący podmiotem przekonań i pragnień musi cechować się wewnętrzną subosobową organizacją umożliwiającą mu posiadanie odpowiednich własności dyspozycyjnych – tych mianowicie, które są konstytutywne dla żywienia postaw propozycjonalnych (por. Bechtel, Abrahamsen 1993). Innymi słowy, postawy propozycjonalne jako własności dyspozycyjne muszą mieć z konieczności jakąś *bazę kategorialną* (por. Choi, Fara 2012). Bazę taką będzie zawsze stanowić subosobowa architektura systemu poznawczego: zorganizowane, działające komponenty składających się na ten system mechanizmów. Jednak o naturze tej subosobowej bazy kategorialnej postaw propozycjonalnych nie chcę tu zakładać nic więcej ponad to, że stoi ona u podstaw określonych dyspozycji. Z przyjmowanej tu perspektywy nawet mieszkaniec Marsa radykalnie różniący się od istot ludzkich na poziomie bazy kategorialnej – czyli subosobowej architektury – mógłby być podmiotem przekonań i pragnień, o ile miałby on odpowiednie „powierzchowne" dyspozycje (por. Schwitzgebel 2010b). Zarówno człowiek, jak i Marsjanin potrzebują jakiejś subosobowej maszynerii, umożliwiającej im posiadanie określonych własności dyspozycyjnych. Przypisując im postawy propozycjonalne, nie przesądzamy jednak, jaka jest ta maszyneria.

poprawnymi (tezę taką można by jednak zapewne przypisać McDowellowi – por. 1994). Matteo Colombo (2013) argumentuje za tezą, że niektóre zjawiska osobowe – na przykład pewne rodzaje uzależnień – nie mogą być optymalnie czy do końca poprawnie wyjaśnione konstytutywnie (tylko) na poziomie osobowym. Zjawiska te są bowiem (współ)konstytuowane przez fakty z poziomu subosobowego. Innymi słowy, Colombo twierdzi, iż niektóre zjawiska z poziomu osobowego domagają się głębokich wyjaśnień konstytutywnych. Teza ta pozostaje niesprzeczna z bronionymi tu twierdzeniami dopóty, dopóki do klasy fenomenów osobowych domagających się takich głębokich wyjaśnień nie zaliczymy postaw propozycjonalnych.

Wspomnienie o mieszkańcach Marsa naturalnie kieruje nas ku drugiemu minimalnemu warunkowi architekturalnemu, który powinien być spełniony przez posiadaczy przekonań i pragnień. Peacocke (1983: 205) sformułował eksperyment myślowy, w którym pewna istota (nazywa po prostu „Ciałem"), morfologicznie i behawioralnie nieodróżnialna od normalnego człowieka, okazuje się w istocie marionetką sterowaną na odległość przez Marsjanina. Działania takiej marionetki pozornie można interpretować, przewidywać i wyjaśniać za pomocą psychologii potocznej. Jak się jednak wydaje, odkrywszy, że mamy do czynienia z marionetką, której działaniami nie kierują żadne („własne") wewnętrzne procesy neuronalne albo obliczeniowe, powinniśmy odmówić tej istocie statusu podmiotu przekonań czy pragnień. Choć ten eksperyment myślowy miał pierwotnie stanowić argument przeciwko teorii systemów intencjonalnych Dennetta, proponuję wyciągnąć z niego inny wniosek. Jak się wydaje, przykład marsjańskiej marionetki każe nam po prostu nałożyć na podmioty postaw propozycjonalnych kolejny minimalny warunek architekturalny. Jak zauważają Dennett (2008) oraz Henderson i Horgan (2004), eksperyment myślowy Peacocke'a pokazuje jedynie, że warunkiem koniecznym zasadnego uznawania danej istoty za podmiot postaw propozycjonalnych jest to, by jej zachowania były kierowane *własnymi* stanami i procesami mentalnymi[40]. Mówiąc językiem Dennetta (2003), byty „naprawdę przekonane" (*true believers*) muszą być *autonomiczne*. Systemy poznawcze stanowiące podmioty przekonań i pragnień nie mogą być sterowane zewnętrznie; muszą być kontrolowane przez własne wewnętrzne mechanizmy su-

[40] Dennett (2008) zauważa nawet, że przykład ten można zmodyfikować w sposób czyniący „marsjańską marionetkę" pełnoprawnym posiadaczem postaw propozycjonalnych. Specyfika tej istoty polegałaby na tym, że mechanizmy subosobowe, które w normalnych okolicznościach znajdują się wewnątrz zamieszkującego Ziemię organizmu (na ogół wewnątrz jego czaszki), w tym przypadku znajdowałyby się na Marsie. Marsjańska marionetka okazałaby się wtedy po prostu niezwykle „rozproszoną" przestrzennie osobą (systemem poznawczym), której mózg znajduje się na innej planecie niż kontrolowane przezeń ciało.

bosobowe[41]. Taki warunek nie mówi nam jednak nic o tym, jakie to mechanizmy, stąd kwalifikuję go jako *minimalny* warunek architekturalny bycia podmiotem postaw propozycjonalnych.

Po czwarte wreszcie, warto postawić pytanie o relację między tezą o postawach propozycjonalnych jako własnościach wyższego rzędu a problemem naturalizacji intencjonalności. Czy przyjęcie, iż postawy propozycjonalne to własności systemowe, jest do pogodzenia z tezą, że intencjonalność może być znaturalizowana? Należy tu poruszyć kilka kwestii. Zacznijmy od rozróżnienia, które zostało pominięte w przedstawionej w rozdziale 1 (podrozdział 1.2) charakterystyce problemu naturalizacji intencjonalności, a które powinno zostać teraz wprowadzone, jeśli chcemy posunąć się naprzód w rozważaniach. Chodzi o odróżnienie *celu* przyświecającego staraniom zmierzającym do znaturalizowania intencjonalności od *środków* służących jego realizacji. Cel projektu naturalizacji jest bardzo ogólny. Polega on na uzgodnieniu naukowego obrazu świata z twierdzeniem o istnieniu przekonań, pragnień, intencji i innych osobowych stanów intencjonalnych. Z kolei powszechnie uznawany filozoficzny środek służący do jego realizacji polega na sformułowaniu nieintencjonalnych warunków wystarczających dla bycia postawą propozycjonalną o określonej treści, warunków, które mogłyby zostać spełnione przez jakieś stany czy struktury o charakterze subosobowym.

Cel projektu naturalizacji i wymieniony środek realizacji tego celu nie są na ogół od siebie odróżniane w literaturze. Nie bierze się zazwyczaj pod uwagę możliwości, że do naturalizacji intencjonalnych stanów osobowych moglibyśmy dojść inaczej, niż przez wskazanie nieintencjonalnych warunków wystarczających do bycia przekonaniem czy pragnieniem. Skuteczną naturalizację utożsamia się ze wskazaniem takich warunków. Wśród filozofów coraz bardziej widać jednak zwątpienie, czy tak rozumiany projekt naturalizacji intencjonalności ma jakiekolwiek szanse powodzenia (Godfrey-Smith

[41] Nie ulega wątpliwości, że takie rozwiązanie rodzi inny problem filozoficzny: co to znaczy, że pewien system jest sterowany przez „własne" mechanizmy subosobowe? Na czym polega tego rodzaju autonomia? Gdzie dokładnie leży linia demarkacyjna oddzielająca systemy autonomiczne od sterowanych zewnętrznie „marionetek"? Pozostawiam to zagadnienie na boku.

2004; Lycan 2008). Od dawna są także wyrażane wątpliwości dotyczące sensowności takiego projektu (Stich 1992; Baker 1995). Jak dobitnie stwierdza Stich:

> Na czym polega bycie [fonemem] /p/ lub /b/? Jeśli pragniesz odpowiedzi naturalistycznej, takiej, która podaje konieczne i wystarczające warunki w kategoriach fizyki lub biologii, to obawiam się, że się rozczarujesz. Mimo wielu lat wyrafinowanych badań, nie istnieje naturalistyczna odpowiedź na to pytanie. [...] To samo mogłoby zostać powiedziane o innych pojęciach cechujących się niekwestionowaną użytecznością naukową. W etologii naczelnych nie występuje naturalistyczna koncepcja iskania [*grooming behavior*]. [...] Jednakże negowanie istnienia iskania ze względu na to, że nie możemy go zdefiniować za pomocą języka fizyki i biologii, z pewnością byłoby po prostu przekorne. Odpowiednio wytrenowani obserwatorzy potrafią rozpoznawać iskanie (albo fonemy) w sposób wysoce intersubiektywnie wiarygodny. A to, jak sądzę, więcej niż wystarczy do tego, aby uczynić owo pojęcie empirycznie wartościowym. Wymaganie czegoś więcej – w szczególności, aby omawiane kategorie zostały „znaturalizowane" – wydaje się bezzasadne i zabawne. Sytuacja reprezentacji mentalnych wygląda całkiem podobnie (Stich 1992: 258).

Stich twierdzi zatem, że istnieje ogrom zjawisk, które zarazem (1) *prima facie* wydają się nie generować dla naturalisty żadnych problemów oraz (2) nie poddają się naturalistycznej „analizie" w kategoriach dostarczanych przez nauki szczegółowe. Czy nie powinniśmy uznać, że do grupy tej należą także postawy propozycjonalne? Przyjmijmy wraz ze Stichem, że tak rzeczywiście jest: intencjonalność nie może być znaturalizowana w przyjmowanym zazwyczaj rozumieniu „naturalizacji". Zastanówmy się, dokąd mogłaby zaprowadzić taka teza.

Klasycznie pojęty projekt naturalizacji intencjonalności – ten, do którego odwołuje się Stich w przytoczonym cytacie – moglibyśmy sklasyfikować jako zmierzający do *twardej* naturalizacji. Kwalifikacja ta opiera się na fakcie, że zarówno środek realizacji, jak i warunek powodzenia tego projektu są bardzo mocne czy restrykcyjne.

Wyznaczenie naturalistycznych, nieintencjonalnych warunków wystarczających do bycia stanem intencjonalnym to po prostu bardzo mocny wymóg. Chcę zasugerować, że ogólny cel projektu naturalizacji intencjonalności można zrealizować za pomocą mniej kłopotliwych i restrykcyjnych środków. Nazwijmy naturalizację intencjonalności za pomocą takich skromniejszych środków „*miękką* naturalizacją".

Na czym miałaby jednak polegać miękka naturalizacja intencjonalności? W tym miejscu pojawia się rzecz jasna rozległe pole do dyskusji. Biorąc pod uwagę słowa Sticha, moglibyśmy na przykład uznać, iż miękka naturalizacja czegoś wymaga pokazania, że to coś jest „użyteczne naukowo". Na przykład moglibyśmy twierdzić, że przekonania i pragnienia (pojęcia przekonań i pragnień) podlegają naturalizacji dlatego, iż są użyteczne dla socjologa badającego postawy polityczne czy też dla psychologa poznawczego zajmującego się badaniem wpływu pragnień na percepcję (percepcję „życzeniową"). Pytanie odwołujące się do użyteczności naukowej cechuje się jednak brakiem precyzji. Czy pojęcie krzesła jest niekompatybilne z naukowym obrazem świata (nienaturalne/nienaturalizowalne) tylko dlatego, że służy naukowcom w celach życiowo-praktycznych, a nie w celach badawczych? Czy może pojęcie to zostanie znaturalizowane wtedy, gdy pewien historyk wykorzysta je naukowo przy pisaniu książki o przemianach w estetyce mebli w XVIII-wiecznej Europie? Czy użyteczne naukowo są tylko kategorie z zakresu nauk szczegółowych, czy także te z zakresu ekonomii, politologii, literaturoznawstwa, historii sztuki albo sinologii? Podejrzewam, że próżno szukać jednoznacznie uzasadnionych i niearbitralnych odpowiedzi na tego rodzaju pytania.

Pragnę zaproponować inną strategię miękkiej naturalizacji, taką, która wydaje się bardziej precyzyjna i owocna, niż strategia odwołująca się do „użyteczności naukowej". Za moją propozycją stoi idea, że naturalizacja nie wymaga z konieczności *inkorporowania* tego, co naturalizowane, w obręb nauki. Sprzeciwiam się więc „imperialistycznej" wizji naturalizmu, uznającej za wartościowościowe czy prawomocne tylko te kategorie, które bezpośrednio wchodzą w obręb nauk szczegółowych (por. Żegleń 2003). Sądzę, że naturalizując

coś, powinniśmy wymagać jedynie, aby to, co naturalizowane, nie było *nienaturalne*, nie wykraczało „poza" czy „ponad" świat fizyczny, poznawany (poznawalny) metodami naukowymi[42]. Przyjęcie istnienia tego czegoś powinno być zgodne z naukowym obrazem świata, jednak nie ma konieczności, aby to, co podlega naturalizacji, jako takie stanowiło element tego obrazu. Obraz naukowy nie ma pochłonąć obrazu manifestującego się, a jedynie stanowić część jednolitej z nim, niesprzecznej, synoptycznej wizji. Zgodnie zatem z moją propozycją dowolne *W*[43] podlega naturalizacji, o ile:

(N1) Nasza koncepcja tego, czym jest *W*, nie implikuje, że przyjęcie istnienia *W* jest niespójne z tezą o globalnej superwenienccji na fundamentalnych własnościach fizycznych[44].
(N2) Przynajmniej potencjalnie można pokazać (na przykład za pomocą eksperymentu), że *W* jest przyczynowo efektywne w sposób obserwowalny czy wykrywalny dla istot ludzkich.

[42] Warto rozjaśnić, jak rozumiem tu opozycję naturalne–nienaturalne. Może być ona bowiem odczytana przez pryzmat przeciwstawienia tego, co biologiczne, temu, co kulturowe czy artefaktualne. Przy takiej interpretacji naturalna byłaby mrówka lub wątroba, a nienaturalne – wszelkie artefakty kulturowe, takie jak książka, krzesło albo komputer. Nie o taki sposób rozumienia tego, co naturalne i nienaturalne mi chodzi. „Nienaturalne" w przyjętym tu znaczeniu będą wszelkie byty „ponadnaturalne" czy niefizyczne; to znaczy takie, które nie mogą wchodzić lub nie wchodzą w skład świata fizycznego, jakim jest on opisywany przez nauki szczegółowe. Wszelkie inne byty – wliczając w to artefakty kulturowe – mogą zostać uznane za naturalne w przyjmowanym tu sensie.

[43] Domyślnie uznaję, że *W* jest własnością. Wydaje się jednak, iż nic nie stoi na przeszkodzie, by założyć, że lista kandydatów do naturalizacji jest ontologicznie heterogeniczna i może obejmować także obiekty, procesy czy nawet zdarzenia albo stany rzeczy.

[44] Teza ta głosi: Z konieczności każdy świat możliwy, stanowiący dokładną kopię świata aktualnego pod względem fundamentalnych własności fizycznych (i niezawierający żadnych dodatkowych niefizycznych składników), będzie zarazem stanowić dokładną kopię świata aktualnego *simpliciter* (pod względem własności biologicznych, ekonomicznych, intencjonalnych, fenomenalnych, geologicznych, społeczno-kulturowych, estetycznych i tak dalej).

Jednocześnie proponuję uznać, że:

> **(N3)** Dla dowolnego *W*, jeśli *W* zostało znaturalizowane, to pojęcie *P* tego *W* zostało także znaturalizowane.

Warunek (N1) wyraża ideę, że przedmiot naturalizacji nie powinien okazać się niefizyczny. Przyjęcie istnienia *W* powinno być spójne z fizykalizmem, definiowanym za pomocą pojęcia globalnej superweniencji na fundamentalnych własnościach świata fizycznego (por.: Baker 1995: 213–217; Braddon-Mitchell, Jackson 2007: 21––35). Ma to oddawać sprawiedliwość intuicyjnej idei, że byty czy własności naturalne to takie, które nie są „ponadnaturalne", a zatem niefizyczne czy wykraczające poza świat fizyczny. Warunek (N2) jest inspirowany Iana Hackinga (1983) „realizmem bytowym" (*entity realism*). Zawieszam tu osąd w kwestii, czy (N2) dostarcza, jak chciałby Hacking, dobrego kryterium realności czegoś. Przyjmuję jednak za *prima facie* wiarygodne twierdzenie, że znaturalizowany (naturalny) byt powinien przynajmniej potencjalnie wpływać przyczynowo na świat – w sposób, który będzie wykrywalny dla istot ludzkich. Wreszcie (N3) dotyczy nie tyle naturalizowanych własności czy obiektów, ile raczej naszych pojęć tych własności czy obiektów. Zgodnie z (N3) pojęcie jest znaturalizowane, o ile jest znaturalizowane to, co stanowi desygnat tego pojęcia.

Postuluję, że (N1), (N2) i (N3) w znaczącym stopniu oddają sprawiedliwość funkcji, jaką pojęcie naturalizacji pełni, lub mogłoby owocnie pełnić, w dyskursie filozoficznym. Stawiam hipotezę, że warunki (N1) i (N2) będą spełniać tylko takie *W*, które *prima facie* uznalibyśmy za nieproblematyczne dla naturalistów (na przykład fonemy i iskanie), a nie spełniają ich takie byty czy własności, które są *prima facie* problematyczne w świetle naturalizmu (niematerialne dusze, pola morfogenetyczne czy flogiston)[45]. Są to także warunki

[45] Ktoś może podnieść wątpliwość, czy nakładanie na *W* zarówno warunku (N1), jak i (N2) nie jest redundantne. Czy nie jest tak, że dla dowolnego *W*, *W* spełnia warunek (N1) wtedy i tylko wtedy, gdy *W* spełnia warunek (N2)? Odpowiedź jest negatywna. Za *W* spełniające warunek (N1), lecz nie (N2), moglibyśmy uznać chociażby: (1) pewne nieefektywne przyczynowo własności relacyjne,

dużo mniej restrykcyjne, niż warunek dostarczenia definicji czegoś (lub choćby listy warunków wystarczających do bycia tym czymś) w kategoriach czysto fizykalnych albo biologicznych. Dlatego mówienie tu o „miękkiej" naturalizacji wydaje się uzasadnione. Nie ulega wątpliwości, że tego rodzaju koncepcja naturalizacji wymaga dopracowania i rozszerzenia, jednak proponuję ją jako dobry szkic roboczy, zasługujący przynajmniej na prowizoryczną akceptację.

Jakie znaczenie mają powyższe rozważania dla prezentowanej tu koncepcji? Otóż twierdzę, że w świetle teorii postaw propozycjonalnych, którą chcę tu zarysować, te ostatnie rzeczywiście podlegają miękkiej naturalizacji. Uważam za nieproblematyczne twierdzenie, że własności dyspozycyjne systemów poznawczych superweniują globalnie na fundamentalnych własnościach fizycznych. Z przyjmowanego tu punktu widzenia postawy propozycjonalne są własnościami dyspozycyjnymi (wiązkami własności dyspozycyjnych) systemów poznawczych. Zatem przekonania i pragnienia spełniają warunek (N1)[46]. Z konieczności każda kopia świata aktualnego pod

takie jak własność znajdowania się na lewym brzegu Wisły w Toruniu albo własność bycia podziwianym przez małżonkę/małżonka (por. Shoemaker 1980); (2) niektóre byty postulowane przez fałszywe teorie naukowe, takie jak flogiston czy pole morfogenetyczne. Kandydatami na *W* spełniające warunek (N2), lecz nie (N1), mogłyby być z kolei istoty boskie albo dusze kartezjańskie – niefizyczne, lecz przyczynowo efektywne w sposób obserwowalny czy wykrywalny dla ludzi.

[46] Temu twierdzeniu można by postawić jeden zasadniczy zarzut. Zgodnie z teorią, która zostanie przedstawiona w sekcji 5.3.2, bycie podmiotem postaw propozycjonalnych polega między innymi na posiadaniu dyspozycji do znajdowania się w określonych stanach fenomenalnych (egzemplifikowania określonych własności fenomenalnych w określonych okolicznościach). W prowadzonej współcześnie debacie na temat świadomości fenomenalnej istnieją bardzo wpływowe stanowiska głoszące, że własności fenomenalne nie superweniują globalnie na fundamentalnych własnościach fizycznych (por. Chalmers 2010). Jeśli zaakceptujemy taką pozycję, a jednocześnie uznajemy własności fenomenalne za współkonstytutywne dla postaw propozycjonalnych, to możliwość naturalizacji tych ostatnich – lub przynajmniej możliwość ich *kompletnej* naturalizacji – może zostać podana w wątpliwość. Mówiąc wprost, powstaje problem, czy za posiadaczy przekonania, że *p*, mogą zostać uznane *zombie* pozbawione dyspozycji fenomenalnych współkonstytuujących posiadanie przekonania, że *p*. Niestety dostarczenie pełnej i zadowalającej odpowiedzi na tę

względem dystrybucji fundamentalnych własności fizycznych (i niezawierająca nic poza nimi) będzie stanowić kopię świata aktualnego pod względem „dystrybucji" przekonań i pragnień. Ponadto, jak postaram się pokazać w sekcji 5.3.3, postawy propozycjonalne są przyczynowo efektywne. Nie tylko można nimi manipulować i zmieniać przez to świat w obserwowalny czy wykrywalny sposób, ale ich przyczynowa efektywność po prostu polega na tym, że można nimi tak manipulować. Również warunek (N2) zostaje zatem spełniony. Skoro zaś (N1) i (N2) są spełnione dla postaw propozycjonalnych, to w zgodzie z (N3) także potoczne pojęcia postaw propozycjonalnych mogą zostać uznane za (miękko) znaturalizowane. Skoro zaś wszystkie trzy warunki są spełnione, można powiedzieć, że postawy propozycjonalne mogą wchodzić w obręb „synoptycznej wizji" postulowanej przez Wilfrieda Sellarsa. Są one – w szerokim sensie – naturalne, czy też stanowią część świata naturalnego.

5.3.2. Postawy propozycjonalne jako własności dyspozycyjne systemów poznawczych

W świetle dyspozycyjnego podejścia do postaw propozycjonalnych naturę przekonań, pragnień czy intencji możemy rozpatrywać przez analogię do cech osobowości (Schwitzgebel 2002). Weźmy choćby pod uwagę taką cechę, jak brawurowość. Na czym polega bycie osobą brawurową? Można twierdzić, że opiera się ono konstytutywnie na posiadaniu szeregu dyspozycji do działania w sposób, który uznalibyśmy intuicyjnie za brawurowy. Brawura jest sposobem, w jaki ktoś żyje i porusza się w świecie naturalnym i społeczno-kulturowym. Manifestuje się ona w (często spontanicznym) po-

wątpliwość zdecydowanie wykracza poza ramy problemowe tej pracy. Kwestia miejsca świadomości fenomenalnej w świecie fizycznym to temat rozległy i złożony. Zagłębianie się w to zagadnienie oddaliłoby nas od sedna prowadzonych tu rozważań. Zaznaczę jedynie, że przyjmuję tu bez argumentu, iż własności fenomenalne superweniują globalnie na fundamentalnych własnościach fizycznych (z konieczności każdy fizyczny duplikat świata aktualnego stanowiłby fenomenalny duplikat tego świata). Osobowe stany intencjonalne podlegają więc naturalizacji nawet przy założeniu, iż znajdowanie się w tych stanach wymaga posiadania określonych dyspozycji fenomenalnych.

dejmowaniu działań niepotrzebnie i „nadmiarowo" ryzykownych, szczególnie kiedy działania te są widoczne dla innych ludzi. W odpowiednich okolicznościach osoba brawurowa będzie podejmować – z prawdopodobieństwem wyższym niż osoba, która brawurowa nie jest – takie działania, jak: uprawianie sportów ekstremalnych, zjadanie egzotycznych i potencjalnie szkodliwych potraw, wyprzedzanie „na trzeciego", prowokowanie agresywnych i niebezpiecznych zwierząt. Innymi słowy, osoba brawurowa posiada szereg własności dyspozycyjnych: dyspozycji do podejmowania określonych działań w określonych okolicznościach. Brawurowość opiera się na pewnych systematycznych prawidłowościach czy wzorcach działania przejawianych przez dane osoby. W podobny sposób moglibyśmy scharakteryzować takie cechy, jak chociażby tchórzostwo, ekstrawertyzm, ufność czy nieśmiałość.

Zgodnie z dyspozycyjnym podejściem do postaw propozycjonalnych natura osobowych stanów intencjonalnych jest podobna do natury cech osobowości[47]. Przekonania i pragnienia z pewnością

[47] Ktoś mógłby zarzucić takiemu stanowisku, że akceptując dyspozycyjną koncepcję przekonań i pragnień, ujmuje się naturę postaw propozycjonalnych w sposób bardzo zbliżony do tego, jak są rozumiane w kognitywistyce reprezentacje ukryte. Zarówno jedne, jak i drugie – czyli postawy propozycjonalne i reprezentacje ukryte – są wszakże własnościami dyspozycyjnymi systemu poznawczego (lub ściśle z tymi własnościami powiązane). W rozdziale 3 (sekcja 3.3.2) została jednak przytoczona i przyjęta za konkluzywną Ramseya krytyka pojęcia reprezentacji ukrytych. Jak można zatem utrzymywać, że na podobne zarzuty nie jest wystawiona teoria, zgodnie z którą postawy propozycjonalne to własności dyspozycyjne systemu poznawczego? Wątpliwość ta okazuje się bezpodstawna. Między pojęciem reprezentacji ukrytych a przyjmowanym tu rozumieniem postaw propozycjonalnych zachodzą zasadnicze różnice. Pojęcie reprezentacji ukrytych miało (1) desygnować struktury odgrywające rolę w mechanistycznych (subosobowych) wyjaśnieniach zjawisk poznawczych; (2) desygnować struktury zawarte wewnątrz mechanistycznej (subosobowej) architektury systemu poznawczego (reprezentacje wewnętrzne). Przestawiona w rozdziale 3 krytyka pojęcia reprezentacji ukrytych była prowadzona właśnie w kontekście tych dwóch oczekiwań czy wymogów. Mówiąc o postawach propozycjonalnych jako własnościach systemowych, pełniących rolę w wyjaśnieniach osobowych, odrzucam tu zarówno wymóg (1), jak i (2). Po pierwsze, otwarcie porzucam tu ideę, jakoby wyjaśnienia odwołujące się do postaw propozycjonalnych miały być mechanistyczne. Rola eksplanacyjna przekonań

różnią się od cech osobowości pod pewnymi względami; są one zapewne mniej od nich trwałe, w większym stopniu podlegają świadomej kontroli, a także są bardziej zawężone, jeśli chodzi o spektrum okoliczności, w których mogą się manifestować. Jednak podobnie jak posiadanie cech osobowości, tak posiadanie postaw propozycjonalnych polega na „życiu w określony sposób" (Schwitzgebel 2013). Być przekonanym, że *p*, to działać zgodnie z pewnym wzorcem. Bardziej precyzyjnie – przy takim ujęciu bycie podmiotem postawy propozycjonalnej polega na posiadaniu określonej własności dyspozycyjnej czy szeregu takich własności. Zgodnie z podejściem dyspozycyjnym osoba czy system poznawczy *S* jest podmiotem postawy propozycjonalnej o określonej treści, jeśli można o tej osobie/ /systemie prawdziwie orzec szereg okresów kontrfaktycznych o postaci: „Jeśli wystąpi okoliczność *O*, to *S* (prawdopodobnie) znajdzie się w stanie *Z*" (Baker 1995: 153–192; Schwitzgebel 2002, 2013). Stan *Z* stanowi w takiej sytuacji manifestację własności dyspozycyjnej, natomiast *O* możemy nazwać okolicznością „wywołującą" manifestację (Schwitzgebel 2002). Przypisując osobie przekonanie, że *p*, stwierdzamy, co dzieje lub działoby się z tą osobą (jak działałaby ona), gdyby wystąpiły określone okoliczności.

W literaturze filozoficznej występuje kilka teorii postaw propozycjonalnych, które można by uznać za mieszczące się w ramach podejścia dyspozycyjnego (por.: Ryle 1970; Audi 1994; Baker 1995: 153–192; Schwitzgebel 2002, 2010b, 2013; Matthews 2007: 123–256; Ratcliffe 2007: 205–211). Podejście to na ogół jest jednak przede wszystkim (lub nawet jedynie) kojarzone z behawioryzmem analitycznym Gilberta Ryle'a[48]. Podejście dyspozycyjne w odmianie bro-

---

i pragnień zasadniczo różni się od roli subosobowych reprezentacji, którym były poświęcone rozdziały 2 i 3. Po drugie, rezygnuję z idei, jakoby przekonania i pragnienia miały być identyczne z jakimiś elementami subosobowej architektury poznania. Moim celem nie jest obrona tezy, że przekonania i pragnienia są „ukryte" wewnątrz mechanistycznej struktury systemu poznawczego czy „rozproszone" po niej. Twierdzę raczej, że postawy propozycjonalne „znajdują" się na innym, wyższym poziomie organizacji tego systemu.

[48] Należy zaznaczyć, że idee Ryle'a były antycypowane przez „późnego" Wittgensteina (2000) oraz niektórych neopozytywistów (por. chociażby Roberta

nionej przez tego autora (1) miało opierać się na podaniu *pojęciowej analizy* (osobowych) predykatów mentalnych w kategoriach dyspozycyjnych (czyli na zanalizowaniu tych predykatów za pomocą okresów kontrfaktycznych wyrażających różne własności dyspozycyjne podmiotu); (2) analiza taka miała być *redukcyjna*, to znaczy analizans nie powinien był zawierać predykatów mentalnych; (3) dyspozycje, za pomocą których były analizowane predykaty mentalne, miały być czysto *behawioralnymi* dyspozycjami do wykonywania określonych ruchów fizycznych czy sekwencji ruchów fizycznych (manifestacji) w określonych okolicznościach (wywołujących manifestacje)[49].

Propozycja Ryle'a jest współcześnie uważana za konkluzywnie obaloną. Przeciwko behawioryzmowi analitycznemu można bowiem wytoczyć bardzo silne argumenty. Przytoczmy kilka z nich. Po pierwsze, nie sposób dostarczyć „behawiorystyczną" analizę jakiegokolwiek predykatu mentalnego, która byłaby redukcyjna i unikała kolistości. Na przykład motorycznie identyczny ruch podniesienia ręki może być wyrazem intencji zatrzymania taksówki, intencji znalezienia czegoś w wysoko położonej półce lub intencji przywitania znajomego. To, która z tych opcji będzie poprawna, zależy od okoliczności – nie tylko zewnętrznych, ale także tych związanych z posiadaniem innych stanów mentalnych. Ruch ręki będzie manifestacją intencji zatrzymania taksówki tylko wtedy, gdy osoba wykonująca go *wie*, na czym polega instytucja taksówki, *sądzi*, że przejeżdżający samochód jest taksówką, *pragnie* w danym momencie zatrzymać taksówkę i tak dalej. Przekreśla to możliwość dostarczenia takiej dyspozycyjnej analizy predykatów mentalnych, która sama nie odwołuje się do innych predykatów mentalnych. Po drugie, nie-

Poczobuta – 2009: 138–141 – krytyczne omówienie neopozytywistycznego behawioryzmu analitycznego w sformułowaniu Rudolfa Carnapa).

[49] Warto mieć tu na uwadze pewne egzegetyczne obserwacje Erica Schwitzgebela (2002), który sugeruje, że tego rodzaju „standardowa" interpretacja zamierzeń Ryle'a może być nadmiernie uproszczona. Schwitzgebel pokazuje, że u Ryle'a można znaleźć fragmenty, w których sugeruje on możliwość analizowania predykatów mentalnych nie tylko za pomocą dyspozycji behawioralnych, ale także dyspozycji do posiadania określonych świadomych doznań.

trudno wyobrazić sobie logicznie lub wręcz nomologicznie możliwe istoty, które intuicyjnie wydają się podmiotami osobowych stanów mentalnych, a które jednak nie mają odpowiednich dyspozycji behawioralnych. Hilary Putnam (1963) przedstawił przykład „super-Spartan", którzy doświadczają bólu, jednak nigdy nie wyrażają go behawioralnie. Galen Strawson (1994: 254–256) wyobraża sobie z kolei istoty zwane „Obserwatorami Pogody" (*Weather Watchers*), które to – ze względu na posiadane doświadczenia fenomenalne – są podmiotami przekonań i pragnień na temat obserwowanych procesów pogodowych, a jednak są konstytutywnie pasywne, czyli z istoty niezdolne do podejmowania, a nawet wyobrażania sobie zachowań, w których mogłyby manifestować się ich postawy propozycjonalne. Po trzecie, jak zauważa Robert Piłat (1999: 87–88), liczba potencjalnych zachowań odpowiadających dowolnej postawie propozycjonalnej jest ogromna. Wyobraźmy sobie chociażby, na jak wiele sposobów – wliczając w to także zachowania językowe – może behawioralnie manifestować się przekonanie, że demokracja to najlepszy, choć daleki od ideału ustrój polityczny. Projekt dostarczenia kompletnej listy dyspozycji behawioralnych, których posiadanie byłoby warunkiem koniecznym i wystarczającym do bycia podmiotem tej czy innej postawy propozycjonalnej, jest kompletnie nierealistyczny i w praktyce niewykonalny (o czym świadczy także fakt, że nikt nigdy takiej analizy nie dokonał).

Powyższe argumenty sprawiają, że dla wielu współczesnych autorów – kojarzących na ogół podejście dyspozycyjne z behawioryzmem analitycznym – traktowanie postaw propozycjonalnych jako własności dyspozycyjnych jest skazane na porażkę. Wszelkie próby „wskrzeszenia" tego rodzaju podejścia mogą wydawać się filozoficznie wątpliwe. Konstatacja taka nie jest jednak uprawniona, ponieważ istnieją alternatywne rozwinięcia podejścia dyspozycyjnego, które unikają mielizn behawioryzmu analitycznego, a jednocześnie mają określone filozoficzne zalety (por.: Audi 1994; Baker 1995: 153–192; Schwitzgebel 2002, 2013; Matthews 2007: 123–256). Teorie te są na ogół zbliżone do siebie. Podobieństwa objawiają się nie tylko w tym, że owe koncepcje utożsamiają żywienie przekonań i pragnień z posiadaniem określonych dyspozycji, ale też w tym, że w podobny

sposób odbiegają one od twierdzeń behawioryzmu analitycznego Ryle'a. Chciałbym skupić się na jednej z tych teorii – tej, którą uważam za najbardziej rozwiniętą i dopracowaną. Chodzi o wspomnianą wcześniej dyspozycyjno-fenomenalną teorię postaw propozycjonalnych Schwitzgebela (2002; 2013)[50].

Zasadniczym elementem koncepcji Schwitzgebela (2002, 2013) jest pojęcie *stereotypu dyspozycyjnego*[51], na który składa się wiązka własności dyspozycyjnych. Wymienione wcześniej dyspozycje związane z brawurowością mogłyby zostać uznane za wchodzące w obręb stereotypu dyspozycyjnego tej cechy osobowości, ponieważ są one naturalnie i intuicyjnie łączone właśnie z nią. Według Schwitzgebela także przekonania, pragnienia czy intencje o określonych treściach posiadają tego rodzaju stereotypy dyspozycyjne. W obręb stereotypu dla postawy propozycjonalnej o określonej treści wchodzą tylko i wyłącznie te dyspozycje, które zwykle uznalibyśmy – my, użytkownicy psychologii potocznej – za charakterystyczne dla tej postawy, te dyspozycje, które naturalnie i intuicyjnie wiążemy z tą postawą. Zgodnie z centralną tezą Schwitzgebela posiadać postawę propozycjonalną o takiej a takiej treści, to posiadać zestaw własności dyspozycyjnych odpowiadający w określonym zakresie stereotypowi dyspozycyjnemu tej postawy. Mówiąc prościej, być podmiotem przekonania, że *p*, to działać – aktualnie lub potencjalnie – w sposób zgodny ze stereotypem dyspozycyjnym przekonania, że *p*[52].

---

[50] Warto wspomnieć, że w ostatniej publikacji Schwitzgebel (2013) rozszerza swoją koncepcję także na inne niż postawy propozycjonalne rodzaje stanów, mianowicie na „postawy reaktywne" (na przykład nienawidzenie, docenianie, przebaczanie) oraz różne inne postawy, jakie można zajmować względem ludzi, zdarzeń czy obiektów (bycie zakochanym w kimś, bycie fanem piłki nożnej, cenienie inteligencji u innych i tym podobne).

[51] Stereotyp pewnego *x* to po prostu zestaw własności naturalnie łączonych z *x*. Stereotypowe dla bycia ptakiem jest chociażby posiadanie zdolności latania, dzioba czy piór. Stereotypy nie są listami warunków koniecznych i wystarczających do tego, by przynależeć do danej kategorii (pozbawiony piór nielot z niewykształconym dziobem także może być ptakiem). Są one jednak tym bardziej trafne, im bardziej posiadanie stereotypowych własności rzeczywiście świadczy o przynależności kategorialnej.

[52] Należy zauważyć, że według Schwitzgebela (2002) stereotypy dyspozycyjne nie mają charakteru jedynie opisowego, ale także normatywny. Podmiot danej

Weźmy pod uwagę konkretną postawę, na przykład przekonanie, że na zewnątrz jest zimno. Bycie podmiotem tego przekonania naturalnie wiążemy z posiadaniem całego szeregu dyspozycji, obejmujących chociażby: dyspozycję do (przy zajściu odpowiednich okoliczności) przytaknięcia na pytanie o to, czy jest zimno; do ubrania ciepłych ubrań przed wyjściem z domu; do zwrócenia uwagi osobie wybierającej się na zewnątrz w krótkich spodenkach; do nabycia przekonań implikowanych przez sąd: „Na zewnątrz jest zimno"; do odczucia zaskoczenia, jeśli po wyjściu okaże się, że panuje upał i tak dalej. Dopóki wymienione dyspozycje mają to do siebie, że bez problemu uznalibyśmy je za charakterystyczne dla osób przekonanych, że jest zimno, dopóty można uznać, że należą one do stereotypu dyspozycyjnego przekonania, że jest zimno. Jeśli ktoś posiada wymienione dyspozycje (i cały szereg innych, niewymienionych), to będzie przekonany, że jest zimno. Żywienie tego przekonania okazuje się bowiem *konstytuowane* przez spełnianie stereotypu dyspozycyjnego tego przekonania (działanie zgodnie z tym stereotypem).

Schwitzgebel (2002) nie twierdzi, że osoby posługujące się psychologią potoczną potrafią *explicite* wymienić dyspozycje składające się na stereotyp tej czy innej postawy propozycjonalnej. Ludzie na ogół nie tworzą stereotypów dyspozycyjnych w sposób świadomy i kontrolowany. Co więcej, nie w każdym przypadku wszyscy użytkownicy psychologii potocznej będą zgadzać się co do tego, ja-

---

postawy propozycjonalnej nie tylko działa w sposób zgodny ze stereotypem tej postawy, ale takie działanie jest od niego na ogół zarazem oczekiwane, wymagane i egzekwowane przez otoczenie społeczne. Kategorie psychologii potocznej nie tylko opisują i wyjaśniają działania, ale także je kształtują. Idea ta dobrze wpisuje się w nowsze podejścia do badania ludzkiego poznania społecznego, gdzie coraz częściej broni się tezy, iż psychologia potoczna nie stanowi przede wszystkim narzędzia czytania umysłów, lecz jest narzędziem kształtowania i „urabiania" umysłów przez otoczenie społeczne, tak by zapewnić grupom ludzkim pewien zakres poznawczej i behawioralnej homogeniczności (Mameli 2001; Zawidzki 2013). Można by w pewnym uproszczeniu powiedzieć, że z takiej perspektywy istnienie przekonań i pragnień nie jest przyczyną tego, że psychologia potoczna okazuje się tak skuteczna eksplanacyjnie i predykcyjnie, lecz stanowi *skutek* procesu kształtowania wzorców ludzkiego myślenia, odczuwania i działania za pomocą psychologii potocznej.

kie dyspozycje wiążą się z daną postawą propozycjonalną. Będą istnieć pewne własności dyspozycyjne, które są centralne dla stereotypu (uznawane przez zdecydowaną większość za stereotypowe), jak również takie, które są bardziej peryferyjne (rzadziej łączone z daną postawą propozycjonalną). Schwitzgebel nie zaprzecza też temu, że w przypadku każdej postawy propozycjonalnej liczba dyspozycji wchodzących w obręb stereotypu tej postawy będzie ogromna (dla dowolnej postawy moglibyśmy takie dyspozycje wymieniać właściwie bez końca). Jednak fakty te nie stanowią dla omawianej teorii większego problemu – właśnie dlatego, że wspiera się ona na pojęciu stereotypu. Celem Schwitzgebela nie jest dokonanie redukcyjnej analizy osobowych kategorii mentalnych i dostarczenie ich definicji. Autor ten nie jest zobligowany do utrzymywania, że osiągalne (albo przynajmniej możliwe) będzie stworzenie listy warunków koniecznych i wystarczających do posiadania przekonania czy pragnienia o określonej treści. Jego teoria nie dostarcza tego rodzaju definicji i nie ma na celu ich dostarczenia. Ta wersja podejścia dyspozycyjnego do postaw propozycjonalnych pozostaje niesprzeczna z twierdzeniem, że tego rodzaju definicje po prostu nie mogą zostać sformułowane. Posiadanie przekonania, że na zewnątrz jest zimno, nie polega na spełnieniu definicji, która mogłaby zostać sformułowana w wyniku skrupulatnej analizy pojęciowej, lecz – na posiadaniu dyspozycyjnego „profilu" zbieżnego ze stereotypem dyspozycyjnym tego przekonania. To ważny punkt, w którym omawiana teoria odbiega od behawioryzmu analitycznego, uwalniając się od problemów tamtej koncepcji.

Schwitzgebel (2002, 2013) wymienia trzy ogólne rodzaje dyspozycji, które mogą wchodzić w skład stereotypu dowolnej postawy propozycjonalnej. Są to (1) dyspozycje *behawioralne*, (2) dyspozycje *poznawcze* oraz (3) dyspozycje *fenomenalne* (por. Baker 1995). Dyspozycje behawioralne to te, do których (wyłącznie) odwoływał się Ryle. Na przykład w obręb stereotypu przekonania, że jest zimno, może wchodzić dyspozycja behawioralna do wykonania (przy zajściu odpowiednich okoliczności) sekwencji ruchów prowadzących do ubrania cieplejszych ubrań. Dyspozycje poznawcze są związane z tendencjami do formowania innych postaw propozycjonalnych.

Na przykład osoba przekonana, że jest zimno, może mieć dyspozycję do uformowania (przy zajściu odpowiednich okoliczności) przekonania, że temperatura na zewnątrz wynosi mniej niż 20 stopni Celsjusza, czy też do uformowania intencji, by ciepło się ubrać przed wyjściem. Wreszcie dyspozycje fenomenalne dotyczą stanów fenomenalnych. Na przykład osoba przekonania, że jest zimno, może odczuć (przy zajściu odpowiednich okoliczności) nieprzyjemny dreszcz na myśl o opuszczeniu domu albo doznać subiektywnego poczucia zdziwienia, kiedy po wyjściu okaże się, że na zewnątrz panuje upał.

Tak bogate ujęcie rodzajów dyspozycji konstytuujących posiadanie postaw propozycjonalnych uwalnia teorię Schwitzgebela od kolejnych problemów, jakimi jest obarczony behawioryzm analityczny. Jak zostało zaznaczone, zasadniczym problemem dla behawioryzmu pozostaje fakt, iż nie jesteśmy w stanie dokonać redukcyjnej analizy predykatów mentalnych – nie sposób analizować predykaty mentalne bez odwoływania się do innych predykatów mentalnych. Jednak dyspozycyjna teoria Schwitzgebela *nie* ma charakteru redukcyjnego. Autor ten zupełnie otwarcie przyznaje, że posiadanie postaw propozycjonalnych może konstytutywnie zależeć od dysponowania *innymi postawami propozycjonalnymi*, i to na dwa sposoby. Po pierwsze, wśród dyspozycji należących do stereotypu dowolnej postawy propozycjonalnej znajdują się również (poznawcze) dyspozycje do nabywania innych postaw propozycjonalnych. Po drugie, Schwitzgebel (2002) poświęca wiele uwagi faktowi, że wszelkie dyspozycje składające się na stereotyp dowolnej postawy manifestują się jedynie *ceteris paribus*. Dyspozycje te będą się manifestować tylko, o ile nie zachodzą okoliczności tła, które w jakiś sposób uniemożliwiają przejawianie owych manifestacji[53]. W obręb klauzuli *ceteris paribus* może

[53] Możemy przyjąć za Schwitzgebelem (2002), że jeśli pewna osoba nie manifestuje dyspozycji związanych z jakimś przekonaniem (czy inną postawą) ze względu na zachodzenie okoliczności wymienionych w ramach klauzuli *ceteris paribus*, to brak takich manifestacji jest „usprawiedliwiony" (*excused*). Osoba usprawiedliwiona z braku manifestacji danej postawy propozycjonalnej nadal może być uznana za podmiot tej postawy. W ten właśnie sposób będzie na przykład możliwe poprawne przypisywanie przekonań i pragnień osobom sparaliżowanym,

wchodzić wystąpienie pewnych warunków środowiskowych czy biologicznych. Osoba przekonana, że jest zimno, nie będzie manifestować dyspozycji behawioralnych związanych z tym przekonaniem, jeśli jest sparaliżowana, jeśli właśnie śpi lub jeśli w zamieszkiwanym przez nią budynku doszło przed chwilą do eksplozji. Jednak klauzula *ceteris paribus* może obejmować także okoliczności związane z posiadaniem jakichś innych stanów intencjonalnych. Osoba przekonana, że jest zimno, nie będzie manifestować dyspozycji do ubrania się ciepło, jeśli nie ma zarazem *intencji*, by wyjść na zewnątrz, jeśli *pragnie* doświadczyć zimna i tak dalej. Fakty te stanowiłyby problem dla koncepcji Schwitzgebela, gdyby jej celem była pojęciowa redukcja kategorii intencjonalnych[54]. Jak trzeba jednak powtórzyć, autor ten odrzuca taki cel. Jego teoria nie jest i nie ma być konceptualną redukcją tego, co intencjonalne (czy mentalne), do tego, co intencjonalne (mentalne) nie jest.

Szerokie ujęcie dyspozycji przez Schwitzgebela pozwala także oddalić te zarzuty kierowane przeciw dyspozycjonizmowi, które odwołują się do możliwości istnienia podmiotów nigdy nie mani-

w których przypadku postawy te mogą nigdy nie objawić się behawioralnie. Jednym z ważnych problemów dla podejścia dyspozycyjnego pozostaje jednak zagadnienie, jak odróżnić usprawiedliwione braki manifestacji od braków manifestacji świadczących o braku przekonania (pragnienia, nadziei, intencji i tak dalej). Gdzie dokładnie kończy się zakres klauzuli *ceteris paribus*? Schwitzgebel twierdzi, że w wielu przypadkach nie będziemy w stanie wyznaczyć takiej granicy w sposób dokładny, jednoznaczny i niezależny od pragmatycznego kontekstu, w którym dokonuje się atrybucja postawy. To, czy należy uznać, że brak manifestacji jest usprawiedliwiony, czy raczej świadczy on o braku danej postawy, ostatecznie będzie często zależeć od wyczucia i praktycznej „znajomości rzeczy" charakteryzującej osoby, które są na co dzień zanurzone w praktyce interpretowania własnych i cudzych działań w kategoriach intencjonalnych.

54 Fakty te stanowiłyby problem także dla realizowanego przeze mnie w tej książce projektu filozoficznego, gdyby jego celem była twarda naturalizacja postaw propozycjonalnych. Jednak tak nie jest. Wystarczy koncepcja, która pozwala na miękką naturalizację osobowych stanów intencjonalnych. Taka miękka naturalizacja – rozumiana w sposób wyrażony w poprzedniej sekcji – nie wymaga jednak podania *nieintencjonalnych* warunków wystarczających do znajdowania się w określonym stanie intencjonalnym. Dlatego dyspozycyjno-fenomenalna teoria Schwitzgebela jest jak najbardziej akceptowalna w kontekście celów teoretycznych prowadzonego tu wywodu.

festujących behawioralnie posiadanych postaw propozycjonalnych. Kluczem do odpowiedzi na tego rodzaju argumenty będzie rola, jaką odgrywają w omawianej teorii dyspozycje fenomenalne i poznawcze. Super-Spartanie Putnama czy Strawsonowscy Obserwatorzy Pogody nie manifestują behawioralnych własności dyspozycyjnych, które są stereotypowe dla posiadanych przez nich stanów intencjonalnych (mentalnych). Jednak manifestują oni odpowiednie dyspozycje poznawcze i fenomenalne. Z tego też powodu mogą oni w znacznym stopniu odpowiadać stosownym stereotypom dyspozycyjnym. Co więcej, niezdolność do behawioralnego manifestowania swoich stanów stanowi okoliczność usprawiedliwiającą ich z braku takich manifestacji (por. przypis 53).

Jak się zatem wydaje, dyspozycyjno-fenomenalna teoria postaw propozycjonalnych wychodzi bez szwanku z konfrontacji z zarzutami uderzającymi w behawioryzm analityczny Ryle'a. Jest to rzecz jasna dobra wiadomość dla każdego filozofa chcącego bronić tezy, że postawy propozycjonalne są dyspozycyjnymi własnościami systemów poznawczych. Oczywiście można zadać pytanie, czy za tego rodzaju koncepcją świadczy coś więcej niż fakt, że jest ona wolna od problemów behawioryzmu? Czy za ujmowaniem postaw propozycjonalnych jako wiązek własności dyspozycyjnych stoją jeszcze jakieś inne, niezależne racje? Otóż Schwitzgebel (2001, 2002, 2010a, 2013) zwraca uwagę, że tego rodzaju koncepcja dobrze radzi sobie z wyjaśnieniem fenomenu, który nazywa on „półprzekonaniami" (*in-between believing*). Zjawisko półprzekonań możemy zilustrować za pomocą następujących przykładów:

> – **Przykład 1.** Henryk jest troskliwym ojcem i gorącym przeciwnikiem spożywania jakichkolwiek nielegalnych środków psychoaktywnych. Od jakiegoś czasu powtarzają się sytuacje, w których jego syn, Jan, wraca wieczorami do domu w stanie wskazującym, że jest pod wpływem marihuany. Są momenty, kiedy Henryk ze smutkiem uznaje, że Jan pali marihuanę. W inne dni Henryk odrzuca to przekonanie i wymyśla scenariusze, które wyjaśniają nocne powroty Jana bez powoływania się na fakt, iż wcześniej konsumował on narkotyk. Jednak kolejnej

nocy Jan znowu wraca do domu w nadzwyczaj dobrym nastroju i z zaczerwienionymi oczyma, przez co Henryk nabiera pewności, że syn palił marihuanę, otwarcie go o to oskarża i nie jest w stanie uwierzyć jego zapewnieniom, że nigdy nie miał on do czynienia z żadnym narkotykiem. Następnego dnia Henryk udaje się na obiad do rodziny swojego brata. W trakcie rozmowy Henryk krytykuje brata i jego małżonkę za nazbyt liberalne strategie wychowania ich córki (regularnej i otwartej palaczki marihuany), a także szczerze i stanowczo twierdzi, że jego własny syn nigdy nawet nie pomyślałby o spróbowaniu jakichkolwiek nielegalnych substancji odurzających. Co tak naprawdę sądzi Jan? Czy jest on przekonany, że jego syn pali marihuanę? (Przykład zaczerpnięty i zaadaptowany z: Schwitzgebel 2002).

– **Przykład 2.** Joanna jest wykładowczynią filozofii. Niedawno przeczytała ona artykuł *Are you living in a computer simulation?* Nicka Bostroma (2003), na podstawie którego uznała, że jej własna świadomość, jak również to, co uznawała do tej pory za świat zewnętrzny, stanowi najprawdopodobniej wynik symulacji przeprowadzanej na superkomputerze przez jakąś technologicznie zaawansowaną cywilizację. Joanna jest gotowa przytaknąć, kiedy ktoś zapyta ją, czy najprawdopodobniej zamieszkuje wirtualną symulację świata, a na prowadzonych przez siebie wykładach z filozofii nowożytnej jest skłonna mówić studentom, że kartezjański demon może być czymś więcej niż tylko bohaterem eksperymentu myślowego. Kiedy jednak przyjrzymy się codziennemu życiu Joanny, okazuje się, że niewiele się w nim zmieniło. Zachowuje się ona względem „wirtualnych" obiektów, zdarzeń czy instytucji społecznych dokładnie tak samo jak wtedy, gdy nie uważała siebie samej i zamieszkiwanego świata za symulację. Gdyby w restauracji dosiadł się do niej mężczyzna utrzymujący, że jest awatarem cywilizacji, która stworzyła zamieszkiwany przez nią wirtualny świat, Joanna (1) byłaby skonsternowana i przestraszona (zamiast odczuć triumf, że miała rację w sprawie istnienia świata zewnętrznego); (2) szybko oddaliłaby się i zgłosiła menedżerowi, że jest nagabywana przez

dziwną, prawdopodobnie chorą psychicznie osobę (zamiast zadać nagabującemu serię pytań o cel symulacji i naturę jej twórców). Czy Joanna naprawdę sądzi, że jest wirtualną świadomością zamieszkującą wirtualny świat?

Schwitzgebel (2001, 2002, 2010a, 2013) wymienia w swoich pracach wiele analogicznych przykładów. Ze zjawiskiem półprzekonań mamy do czynienia wtedy, gdy dysponujemy pewnymi podstawami, by przypisać danej osobie stan przekonaniowy (czy inną postawę propozycjonalną) o określonej treści, jednak pod innymi względami wzorce myślenia, odczuwania i działania tej osoby odbiegają od tego, czego powinniśmy się spodziewać, gdyby rzeczywiście znajdowała się ona w tym stanie. Wydaje się, że w takich sytuacjach nie możemy jednoznacznie orzec, czy dana osoba jest podmiotem określonej postawy propozycjonalnej, czy nie. Według Schwitzgebela (2001, 2002) półprzekonania stanowią zasadniczy problem dla wszystkich teorii utożsamiających stany intencjonalne psychologii potocznej z wewnętrznymi, subosobowymi reprezentacjami (na przykład zdaniami zapisanymi w języku myśli, umieszczonymi w określonej „skrzynce" funkcjonalnej). Inaczej mówiąc, istnienie półprzekonań jest bardzo trudne do pogodzenia z założeniem korespondencji osobowe–subosobowe. Dlaczego? Z punktu widzenia teorii utożsamiających postawy propozycjonalne z subosobowymi reprezentacjami, zawsze powinna istnieć definitywna i jednoznaczna odpowiedź na pytanie o to, czy dana osoba posiada przekonanie, pragnienie czy intencję o określonej treści, czy nie (Schwitzgebel 2001; 2002; 2010a; 2013). Dowolna osoba albo ma „w głowie" określoną subosobową reprezentację, albo nie – a tym samym albo jest podmiotem danej postawy propozycjonalnej, albo nie. Posiadanie przekonania, że *p*, to z takiego punktu widzenia zawsze kwestia o charakterze „wszystko albo nic". Jak jednak pokazują przykłady półprzekonań, jest to zbyt uproszczony i „statyczny" sposób myślenia o naturze osobowych stanów intencjonalnych.

Zjawisko półprzekonań jest jednak zrozumiałe na gruncie podejścia dyspozycyjnego. Zgodnie z dyspozycyjno-fenomenalną koncepcją Schwitzgebela pewna osoba definitywnie posiada daną po-

stawę propozycjonalną, jeśli działa, myśli i odczuwa w sposób, który całkowicie lub w odpowiednim stopniu odpowiada stereotypowi dyspozycyjnemu tej postawy. Jeśli dana osoba w ogóle nie działa, myśli i odczuwa w sposób odpowiadający temu stereotypowi, to definitywnie nie jest podmiotem danej postawy. Jednak z perspektywy teorii Schwitzgebela między tymi dwiema skrajnościami rozciąga się całe spektrum możliwości pośrednich. Półprzekonania to właśnie tego rodzaju sytuacje. Osoby „półprzekonane" odpowiadają stereotypowi dyspozycyjnemu danej postawy na tyle, że w pewnych kontekstach bylibyśmy skłonni jednoznacznie przypisać im tę postawę. Jednak pod innymi względami czy w innych kontekstach odbiegają one od tego stereotypu i zaczynamy mieć wątpliwości, czy nasza atrybucja była poprawna. Nie chodzi o to, że nie istnieją obiektywne, niezależne od obserwatora fakty dotyczące tego, czy ktoś jest podmiotem danego przekonania albo pragnienia. Własności dyspozycyjne posiadane przez ludzi są całkowicie obiektywne. Chodzi raczej o to, że bardzo często nie możemy definitywnie i jednoznacznie rozstrzygnąć, czy zestaw dyspozycji posiadanych przez daną osobę pozwala już scharakteryzować ją jako podmiot postawy o określonej treści, czy nie. Ludzkie „profile" dyspozycyjne są obiektywne, jednak często okazują się także na tyle złożone i heterogeniczne, że nie da się ich jednoznacznie ująć czy skategoryzować za pomocą potocznych pojęć mentalnych. Zgodnie z teorią dyspozycyjno-fenomenalną na tym właśnie polega istota półprzekonań: to wiązki własności dyspozycyjnych, które nie mogą być jednoznacznie i definitywnie sklasyfikowane za pomocą zasobów pojęciowych psychologii potocznej. Podejście dyspozycyjne jest zatem nie tylko jak najbardziej spójne z istnieniem półprzekonań, ale pozwala nam ono także zrozumieć naturę tego fenomenu.

Ostatnia kwestia, jaką należy poruszyć w tej sekcji, to zagadnienie *treści intencjonalnej* postaw propozycjonalnych. Te ostatnie są przecież uznawane za rodzaj reprezentacji mentalnych właśnie dlatego, że posiadają treści. Przekonania i pragnienia są z konieczności o czymś, posiadają warunki prawdziwości czy spełnienia. Co więcej, uznaje się na ogół, że treść osobowych stanów intencjonalnych jest niewywiedziona (Ramsey 2007: 16–18). Przekonanie czy pragnienie

nie zawdzięcza swojej treści temu, że podlega ono interpretacji przez kogoś (lub przez jakiegoś subosobowego konsumenta) jako posiadające tę treść. Treść to jego własność wewnętrzna. Może się wydawać, że zaletę różnych teorii rozwijanych w ramach ZKOS stanowi fakt, iż bezpośrednio podejmują one właśnie zagadnienie treści. Teorie te skupiają się wszakże na wskazaniu warunków, w których pewien stan subosobowy posiada określoną treść intencjonalną, odpowiadającą treści jakiejś postawy propozycjonalnej. Co o treści przekonań i pragnień mówią jednak teorie dyspozycyjne?

Być może najważniejszym źródłem przekonania o tym, że postawy propozycjonalne posiadają treści – czy raczej stanowią *relacje* między podmiotami postaw a określonymi treściami – są fakty dotyczące języka, jakim posługują się użytkownicy psychologii potocznej (Matthews 2007: 97–102, 2011). Zdanie: „Jan sądzi, że pada deszcz", może być interpretowane jako stwierdzające, iż Jan posiada określoną własność relacyjną. Wydaje się ono wyrażać sąd, że zachodzi określona relacja („bycie przekonanym") między Janem a sądem czy treścią „pada deszcz". Innymi słowy, to przyglądając się *językowym narzędziom* służącym użytkownikom psychologii potocznej, możemy dojść do wniosku, że postawy propozycjonalne są relacjami względem treści intencjonalnych. Jednak własności dyspozycyjne nie są własnościami relacyjnymi, lecz monadycznymi (wewnętrznymi, nierelacyjnymi; por. Matthews 2007: 123–256, 2011). Posiadanie dyspozycji do myślenia, działania i odczuwania w sposób zgodny ze stereotypem dyspozycyjnym przekonania, że *p*, nie polega na istnieniu relacji między podmiotem owego przekonania a sądem *p*. Występuje więc oczywiste napięcie między podejściem dyspozycyjnym a tezą, że postawy propozycjonalne są relacjami względem treści intencjonalnych.

Co może zatem zrobić zwolennik podejścia dyspozycyjnego, znalazłszy się w takiej kłopotliwej teoretycznie sytuacji? Nie obędzie się bez znacznej dozy rewizjonizmu. Kluczowe znaczenie dla zwolennika podejścia dyspozycyjnego ma obserwacja, że jest możliwe zachodzenie rozbieżności między logiczną lub gramatyczną postacią zdań wyrażających atrybucje postaw propozycjonalnych a rzeczywistą naturą samych postaw (Matthews 2007: 102–122, 2011).

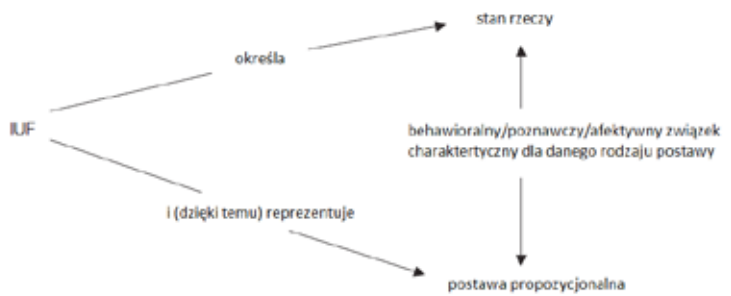
Jeśli dyspozycjonista chce zachować ideę, iż postawy propozycjonalne to dyspozycje czy wiązki dyspozycji, musi odrzucić założenie, że zdania wyrażające atrybucje osobowych stanów intencjonalnych rzeczywiście stwierdzają zachodzenie określonych relacji między podmiotami a sądami/treściami (Matthews 2007: 102–122, 2011).

Nawet jeśli przyjmiemy powyższe rozwiązanie, otwarte pozostają jednak zasadnicze pytania: dlaczego postawy propozycjonalne są opisywane, tak jak gdyby stanowiły relacje względem treści intencjonalnych? Jaka jest zależność między gramatyczną/logiczną postacią zdań wyrażających atrybucje przekonań i pragnień a naturą samych przekonań i pragnień? Te złożone zagadnienia nie mogą być tu systematycznie podjęte. Można jednak zarysować odpowiedź, która została już sformułowana przez Roberta Matthewsa (2007: 123–256; 2011), autora *teoretyczno-pomiarowej* koncepcji postaw propozycjonalnych. W swojej teorii Matthews odwołuje się (między innymi) do następującej idei wyrażonej przez Donalda Davidsona:

> Tak jak dokonując pomiaru wagi, potrzebujemy zestawu bytów cechujących się strukturą, za pomocą której możemy odzwierciedlić relacje między ciężkimi obiektami, tak też przypisując stany przekonaniowe (i inne postawy propozycjonalne), potrzebujemy zestawu bytów powiązanych ze sobą w sposób pozwalający nam na śledzenie istotnych własności różnych stanów psychicznych. [...] nie musimy zakładać, że istnieją takie obiekty [*entities*], jak przekonania. Nie musimy też tworzyć przedmiotów, które by nam służyły jako „przedmioty przekonań", albo jako to, co przedstawia się umysłowi czy mózgowi. Obiekty, do których odwołujemy się, kiedy próbujemy określić stan mentalny, nie muszą odgrywać żadnej roli psychologicznej czy epistemologicznej, tak samo jak liczby nie grają żadnej roli fizycznej [*physical role*] (Davidson 1989: 11).

Weźmy pod uwagę pomiar ciężaru, czyli przykład, do którego odwołuje się Davidson. Kiedy mówimy: „Jan waży 75 kilogramów", nie stwierdzamy, że zachodzi jakaś relacja („ważenie") między Janem a abstrakcyjnym bytem, jakim jest liczba 75. Przypisać obiektowi 75 kilogramów to umieścić go w ramach pewnej skali pomiaru. W omawianym przykładzie skalę pomiaru wyznaczają liczby natu-

ralne. Ważyć 75 kilogramów to posiadać własność fizyczną, dzięki której zajmuje się na stosownej skali miejsce odpowiadające liczbie 75. Mierząc ciężar obiektów, wykorzystujemy izomorfizm lub homomorfizm zachodzący między matematyczną strukturą liczb naturalnych a strukturą potencjalnych wartości, jakie może przyjąć ciężar jako własność fizyczna. Dzięki istnieniu numerycznej skali możemy też wykonać serię rozumowań surogatywnych w sensie Christophera Swoyera (1991). Na przykład za pomocą prostych działań arytmetycznych możemy przewidzieć, jaki ciężar będą miały dwa obiekty, które otrzymamy po tym, jak połączymy ze sobą przedmioty o masie siedmiu i trzech kilogramów, a następnie podzielimy tak powstałe połączenie na dwie połowy o równym ciężarze.



Rysunek 12. Pomiar postaw propozycjonalnych za pomocą skali wyznaczonej przez formy zinterpretowanych wypowiedzi (IUF) w ujęciu Roberta Matthewsa. Źródło: Matthews 2007: 182

Teoretyczno-pomiarowa koncepcja Matthewsa (2007: 123–256; 2011) wspiera się na tezie, że osobowe predykaty intencjonalne – podobnie jak predykaty, za pomocą których określamy ciężar i inne mierzalne własności fizyczne obiektów – stanowią rodzaj predykatów pomiarowych[55]. Przypisując innym przekonania czy pragnie-

[55] Reprezentowanie czegoś za pomocą skali pomiarowej stanowi formę S-reprezentowania tego czegoś: byty podlegające pomiarowi podzielają strukturę

nia o określonych treściach, dokonujemy swoistego *pomiaru* ich stanów mentalnych. Precyzyjniej – mierzy się własności dyspozycyjne podmiotów postaw propozycjonalnych. Według Matthewsa (2007: 150–212) określając rodzaj postawy propozycjonalnej (klasyfikując ją jako przekonanie, pragnienie, oczekiwanie, nadzieję czy intencję), określamy sposób, w jaki podmiot tej postawy będzie działaniowo, poznawczo i afektywnie[56] dysponowany względem pewnego stanu rzeczy. Skalę pomiaru wyznacza zaś to, co autor ten nazywa „formami zinterpretowanych wypowiedzi" (*interpreted utterance forms*, IUF). IUF są odpowiednikami sformułowań występujących zazwyczaj po spójniku „że" w zdaniach wyrażających atrybucje postaw propozycjonalnych (na przykład „pada deszcz" w zdaniu: „Jan sądzi, że pada deszcz"). IUF zachowują wszelkie własności stwierdzeń (*utterances*) występujących po spójniku „że": semantyczne, inferencyjne, syntaktyczne, fonologiczne, pragmatyczne, ortograficzne. Zasadniczą funkcją pomiarową IUF jest określanie stanów rzeczy, względem których podmiot jest behawioralnie, poznawczo i afektywnie ustosunkowany (w sposób charakterystyczny dla danego rodzaju postawy, na przykład przekonania lub pragnienia). Przypisując zatem komuś postawę propozycjonalną o takiej a takiej treści, mierzymy to, jak podmiot tej postawy będzie działaniowo, poznaw-

z bytami składającymi się na skalę pomiaru. Biorąc pod uwagę, że Matthews traktuje osobowe predykaty mentalne jako pomiarowe, otrzymujemy bardzo ciekawą konsekwencję: psychologia potoczna dostarcza nam S-reprezentacji postaw propozycjonalnych (są to jednak S-reprezentacje zewnętrzne – podobne do map, modeli naukowych czy innych pozamentalnych S-reprezentacji – interpretowane przez ludzi, a nie konsumowane przez komponenty subosobowych mechanizmów).

[56] Gdybyśmy chcieli uzgodnić teorię Matthewsa z dyspozycyjno-fenomenalną koncepcją Schwitzgebela, moglibyśmy do tej listy dodać także dyspozycje fenomenalne. Podmiot będzie miał więc nie tylko dyspozycje poznawcze, behawioralne i afektywne względem danego stanu rzeczy, ale także dyspozycje fenomenalne. Przy pewnych założeniach dyspozycje afektywne można nawet uznać za subkategorię dyspozycji fenomenalnych (to jest uznać je za dyspozycje do subiektywnego odczuwania określonych emocji w określonych okolicznościach) albo za hybrydy dyspozycji fenomenalnych i behawioralnych (to jest uznać je za dyspozycje do subiektywnego odczuwania oraz behawioralnego wyrażania określonych emocji w określonych okolicznościach).

czo i afektywnie ustosunkowany względem stanu rzeczy określanego przez dany IUF (por. rysunek 12). Stwierdzając: „Jan sądzi, że pada deszcz", sytuujemy przekonanie Jana na skali wyznaczanej przez IUF i orzekamy, że Jan będzie działał, myślał i emocjonalnie reagował, tak jak gdyby padał deszcz, czy też w sposób biorący pod uwagę, że pada deszcz. Stwierdzając: „Jan pragnie zostać bogatym człowiekiem", orzekamy, że Jan będzie działał, myślał i reagował afektywnie w taki sposób, by ziścił się stan rzeczy polegający na tym, że Jan jest bogaty. W obu tych przypadkach osobowe predykaty intencjonalne służą nam do określenia („zmierzenia") własności dyspozycyjnych Jana, a nie do stwierdzenia faktu, iż znajduje się on w jakieś relacji do abstrakcyjnego bytu, jakim jest sąd czy treść intencjonalna.

Powyższa rekonstrukcja ujmuje zaledwie szkielet teorii Matthewsa. Dla bieżących celów jest ważne jedynie podkreślenie faktu, że podejście dyspozycyjne może obyć się bez postulowania czegoś takiego jak treści czy sądy, w relacje z którymi miałyby wchodzić podmioty postaw propozycjonalnych. W zgodzie z tym, co stwierdza Davidson w przytoczonym wcześniej cytacie, aby móc opisywać osobowe stany intencjonalne oraz myśleć o nich przez przypisywanie im treści, wcale nie musimy postulować, że stany te dosłownie *posiadają* treści albo są *relacjami* względem treści. W ten sposób koncepcje dyspozycyjne nie tyle rozwiązują problem treści intencjonalnej – być może nie ma tu bowiem *de facto* żadnego problemu – co raczej go neutralizują przez odrzucenie tezy, że postawy propozycjonalne są relacjami między podmiotami a sądami (treściami).

5.3.3. Postawy propozycjonalne jako przyczyny

Jak już zostało kilkukrotnie wspomniane, we współczesnej filozofii umysłu niemal konsensualnie uznaje się twierdzenie, że przekonania, pragnienia i inne postawy propozycjonalne są efektywne przyczynowo, a horyzontalne wyjaśnienia działań za pomocą psychologii potocznej to wyjaśnienia przyczynowe. Przyjmowane tu przeze mnie podejście do natury postaw propozycjonalnych może jednak wydawać się niespójne z tezą o przyczynowej efektywności osobowych stanów intencjonalnych. Utrzymuję, że postawy propozycjo-

nalne są własnościami wyższego rzędu (własnościami systemowymi). Co więcej, twierdzę zarazem, że przekonania i pragnienia są własnościami dyspozycyjnymi, a precyzyjniej – wiązkami własności dyspozycyjnych. Jednakże zarówno przeciwko przyczynowej efektywności własności wyższego rzędu, jak i przeciwko przyczynowej efektywności własności dyspozycyjnych – można wysunąć argument z wykluczenia przyczynowego.

Argument z przyczynowego wykluczenia własności wyższego rzędu przyjmuje, w zarysie, postać następującego rozumowania (por. Kim 2002). Załóżmy, że chcemy wyjaśnić przyczynowo zjawisko Z (na przykład wyjęcie parasola z torby przez pewną osobę) za pomocą własności systemowej S (na przykład intencji, by wyjąć parasol), która jest realizowana przez własność niższego rzędu F (na przykład jakąś własność neurofizjologiczną). Czy S mogło przyczynowo wywołać Z? Według zwolenników argumentu z przyczynowego wykluczenia odpowiedź będzie negatywna, ponieważ warunkiem nomologicznie wystarczającym do zajścia Z była egzemplifikacja własności F. Przypisywanie dodatkowo efektywności przyczynowej S przeczyłoby twierdzeniu o niemożliwości „przedeterminowania" (*overdetermination*) skutków, głoszącej, że żaden skutek nie może mieć zarazem dwóch przyczyn, z których każda jest wystarczająca do jego zajścia. W ten sposób S zostaje „wykluczone" przez F jako przyczyna zajścia Z. Cała „moc przyczynowa" zostaje usytuowana we własności F, a własność S – jako własność wyższego rzędu – okazuje się nieefektywna przyczynowo (epifenomenalna).

Analogiczny zarzut można wystosować przeciwko przyczynowej efektywności własności dyspozycyjnych (por.: McKitrick 2005; Choi, Fara 2012). Załóżmy tym razem, że pytamy o to, czy kruchość pewnej szklanki – czyli przysługująca jej własność dyspozycyjna – mogła stanowić przyczynę zbicia jej w pewnym momencie *t*. Zauważmy, że „kandydatem" na przyczynę zbicia szklanki jest także baza kategorialna jej kruchości, czyli własność fizyczna (powiedzmy wiązania chemiczne między atomami, z których jest zbudowana szklanka), na podstawie posiadania której szklanka jest krucha. Co więcej, zwróćmy uwagę na fakt, iż to właśnie posiadanie tej własności fizycznej (oraz pewna okoliczność zewnętrzna, na przykład

kolizja szklanki z podłogą) było w *t* warunkiem nomologicznie wystarczającym do zbicia szklanki. Raz jeszcze, aby uniknąć przedeterminowania skutku, powinniśmy efektywność przyczynową usytuować w bazie kategorialnej (wiązaniach atomowych), a nie własności dyspozycyjnej (kruchości). Baza kategorialna wyklucza przyczynową efektywność własności dyspozycyjnej. W analogiczny sposób można argumentować, że kiedy ktoś wykonuje w momencie *t* działanie, w którym manifestuje się przekonanie, że *p*, to za wykonanie tego działania przyczynowo odpowiada w istocie nie samo przekonanie (jako własność dyspozycyjna), lecz jego baza kategorialna. Co stanowi tę bazę kategorialną? Jest nią wewnętrzna, mechanistyczna architektura czy organizacja mechanizmów składających się na dany system poznawczy, architektura sprawiająca, że system ten dysponuje szeregiem dyspozycji wchodzących w skład stereotypu dyspozycyjnego przekonania, że *p*. Jak się wydaje, lokalizowanie efektywności przyczynowej we własności dyspozycyjnej (przekonaniu), a nie jej subosobowej bazie kategorialnej, jest błędem[57].

[57] Istnieje także inny argument, który mógłby zostać wytoczony przeciwko tezie o przyczynowej efektywności postaw propozycjonalnych jako własności dyspozycyjnych. Zgodnie z tym rozumowaniem relacja między posiadaniem własności dyspozycyjnej a manifestacjami tej własności ma naturę pojęciową czy analityczną (McKitrick 2005; Choi, Fara 2012). Posiadanie przez szklankę dyspozycyjnej własności bycia kruchą nie wywołuje przyczynowo zbicia naczynia (manifestacji), lecz jest z nim powiązane pojęciowo. Posiadanie cechy kruchości *oznacza*, że szklanka zbije się w określonych okolicznościach. Związki pojęciowe nie mogą być związkami przyczynowymi, dlatego własności dyspozycyjne nie mogą wywoływać przyczynowo swoich manifestacji. Analogiczny argument można by wystosować przeciwko przyczynowej efektywności osobowych stanów intencjonalnych (o ile rzecz jasna te są pojmowane dyspozycyjnie). Na przykład można twierdzić, że wybranie numeru w telefonie jest pojęciowo, a nie przyczynowo powiązane z pragnieniem, aby do kogoś zadzwonić. Tego rodzaju argumentacja bez wątpienia uderza w behawioryzm analityczny Ryle'a z jego postulatem przeprowadzenia pojęciowej analizy predykatów mentalnych w kategoriach dyspozycyjnych. Nietrudno jednak zauważyć, w jaki sposób przyjmowana tu za Schwitzgebelem propozycja dotycząca natury postaw propozycjonalnych unika tego problemu. Podejście dyspozycyjne w przyjętej tu odmianie rezygnuje z idei, jakoby można było wyznaczyć dyspozycje, których obecność byłaby konieczna i wystarczająca do posiadania określonej postawy propozycjonalnej. Rezygnuję tu więc z idei, jakoby była możliwa

Zatem argument z wykluczenia przyczynowego wydaje się uderzać w przyjmowaną tu teorię postaw propozycjonalnych – zarówno w twierdzenie o osobowych stanach intencjonalnych jako własnościach wyższego rzędu, jak i w twierdzenie o dyspozycyjnej naturze tych stanów. Pytanie o możliwość uzgodnienia bronionej tu koncepcji z tezą o przyczynowej efektywności przekonań i pragnień moglibyśmy więc zamienić na następujące, bardziej określone: czy można nadać taki sens idei, że postawy propozycjonalne – rozumiane jako systemowe własności dyspozycyjne – są przyczynowo efektywne, by nie narażać się zarazem na argument z przyczynowego wykluczenia? W dalszej części tej sekcji chcę naszkicować tego rodzaju „uodpornioną" na zarzut z przyczynowego wykluczenia koncepcję.

Zacznę odpowiedź na przedstawione wyżej zarzuty od poczynienia ważnej obserwacji dotyczącej argumentu z przyczynowego wykluczenia własności wyższego rzędu. W tym argumencie zasadniczą rolę odgrywa rozróżnienie między własnością wyższego rzędu a realizującą ją własnością niższego rzędu. Jak jednak zauważa Carl Craver (2007: 211–217), ten argument nie może uderzać w tezę o przyczynowej efektywności własności wyższego rzędu, o ile teza ta jest interpretowana przez pryzmat *mechanicyzmu*. Poziomy mechanizmów to bowiem poziomy hierarchicznej kompozycji, a nie realizacji (por. sekcja 2.2.1). Mówienie o efektywności przyczynowej zjawisk z wyższego poziomu (systemowych) w kontekście mechanicyzmu sprowadza się do niekontrowersyjnego twierdzenia, że systemy czy mechanizmy jako całości potrafią zachowywać się w sposób, w jaki nie potrafią działać ich pojedyncze komponenty. W realnych, nieagregatywnych systemach fizycznych nie istnieje na ogół nic, co mogłoby stanowić jakiś odrębny, pojedynczy realizator funkcji systemowej. Na przykład to system poznawczy jako całość, a nie żaden

pojęciowa analiza prowadząca do *zdefiniowania* danego przekonania czy pragnienia w kategoriach dyspozycyjnych. Teza o związku przekonań i pragnień z ich potencjalnymi manifestacjami nie jest tezą semantyczną (por. Baker 1995). Zasadnicza przesłanka omawianego argumentu – czyli twierdzenie o istnieniu pojęciowego/analitycznego związku między postawami propozycjonalnymi (jako własnościami dyspozycyjnymi) a ich manifestacjami – jest zatem fałszywa.

jego komponent, potrafi percypować obiekty albo przypisywać stany umysłowe innym. Analogicznie – to system jako całość, a nie któryś z jego komponentów, może działać na podstawie przekonania (pragnienia, intencji, oczekiwania i tak dalej), że *p*. Z mechanistycznego punktu widzenia twierdzenie o przyczynowej efektywności własności systemowych oznacza po prostu, że całości „mogą więcej" niż ich części.

Zauważmy jednak, że powyższa odpowiedź nie uwalnia nikogo od problemu wykluczenia przyczynowego postaw propozycjonalnych *jako dyspozycji*. Przyjmijmy, iż przekonania i pragnienia są wiązkami własności dyspozycyjnych. Kiedy pytamy o bazę kategorialną przekonania, że *p*, wcale nie musimy szukać jakiegoś pojedynczego wewnętrznego komponentu każdorazowo odpowiadającego przyczynowo za manifestacje dyspozycji należących do stereotypu tego przekonania (stanowiąc w ten sposób realizator tego przekonania). Co innego niż taki pojedynczy komponent może zatem stanowić bazę kategorialną przekonań i pragnień? Jak się wydaje, powinniśmy za takową uznać *globalną*, wewnętrzną organizację mechanistyczną systemu poznawczego (lub przynajmniej jakąś jej rozległą część). Każdy system poznawczy wykazuje określone własności dyspozycyjne dlatego, że dysponuje on specyficzną, złożoną, odpowiednio „skonfigurowaną" mechanistyczną strukturą wewnętrzną. Fakt ten wydaje się niesprzeczny z ustaleniami poczynionymi w poprzednim akapicie. Odpowiednio skoordynowany układ działających komponentów systemu poznawczego jest w stanie „robić" to, czego nie może wykonywać żaden pojedynczy, odrębny komponent. Zasadniczy problem pozostaje zatem otwarty i może być sformułowany w następujący sposób: czy efektywność przyczynowa globalnej, „rozlanej" po systemie poznawczym bazy kategorialnej nie wyklucza efektywności przyczynowej przekonań i pragnień pojmowanych jako wiązki dyspozycji?

Powyższa uwaga pozwala jednak także na rehabilitację argumentu z wykluczenia przyczynowego w sformułowaniu odwołującym się do pojęcia realizacji. Możemy bowiem, jak się wydaje, tak rozwinąć wspomniany termin, aby realizatorami postaw propozycjonalnych mogły być nie tylko wyodrębnione komponenty systemu

poznawczego, ale także stany globalnej wewnętrznej architektury tego systemu. Wprowadźmy pojęcie „Globalnego Stanu Subosobowego" (GSS). GSS to nie tyle komponent czy stan komponentu systemu poznawczego, co raczej stan, w którym jednorazowo znajduje się globalna, zorganizowana mechanistyczna architektura tego systemu. GSS jest wyznaczony przez stany i konfiguracje (sposoby zorganizowania), w jakich znajdują się jednorazowo wszystkie komponenty wszystkich mechanizmów (lub jakaś ich rozległa część) składających się na dany system poznawczy. Załóżmy także, że dwa (egzemplarycznie różne) GSS, w jakich system poznawczy znajduje się w momentach t_1 i t_2, są rodzajowo identyczne, jeśli w t_1 i t_2 wszystkie komponenty wszystkich mechanizmów tego systemu (lub jakaś ich rozległa część) znajdują się w identycznych stanach i konfiguracjach[58].

Wyobraźmy sobie teraz, że Jan wykonuje sekwencję ruchów prowadzących do wyjęcia parasola. Biorąc pod uwagę okoliczności oraz dostępne świadectwa behawioralne, możemy prawidłowo przypisać mu pragnienie, by nie zmoknąć, i sformułować horyzontalne wyjaśnienie przyczynowe na poziomie osobowym: „Jan wyjął parasol, ponieważ pragnął, aby nie zmoknąć". Zajrzyjmy teraz do mózgu Jana, na poziom subosobowy. Przyjmijmy, że Jan (jego system poznawczy) tuż przed zainicjowaniem sekwencji ruchów prowadzących do wyjęcia parasola znalazł się w pewnym GSS, którego wystąpienie było warunkiem nomologicznie wystarczającym do wykonania tej sekwencji ruchów. Można by powiedzieć, że GSS stanowił niezlokalizowany, rozproszony po ośrodkowym układzie nerwowym Jana (czy pewnym jego fragmencie) subosobowy realizator

[58] Pojęcie GSS bez wątpienia wymaga doprecyzowania i rozwinięcia. Na jakim poziomie organizacji systemu poznawczego powinniśmy indywiduować GSS? (Nie chodzi przecież o to, by GSS był indywiduowany z dokładnością sięgającą poziomu aktywności każdego pojedynczego neuronu). Czy każdy GSS może być postrzegany jako punkt w jakiejś „globalnej" przestrzeni stanów opisującej system poznawczy? Czy GSS naprawdę musi obejmować cały system poznawczy, czy tylko jakąś (jaką?) jego część? Mimo tych oraz całego szeregu innych możliwych do podniesienia wątpliwości, mam nadzieję, że ta wstępna, intuicyjna charakterystyka wystarczy na obecne potrzeby.

jego pragnienia. Skoro jednak wystąpienie tego GSS wystarczyło do wyjęcia parasola przez Jana, to jak można utrzymywać tezę o roli przyczynowej jego pragnienia? Pragnienie Jana, aby nie zmoknąć, wydaje się całkowicie epifenomenalne.

Sytuacja komplikuje się dodatkowo, kiedy zwrócimy uwagę na fakt, że zgodnie z przyjmowanym tu podejściem posiadanie tego czy innego przekonania jest konstytuowane przez wzorzec działania (aktualnego i potencjalnego), który pozostaje „widoczny" tylko na poziomie osobowym (na poziomie systemu poznawczego jako całości). Bycie podmiotem postaw propozycjonalnych nie zależy, lub zależy jedynie trywialnie, od faktów z poziomu *subosobowego*. Tak pojmowana mechanistyczna neutralność psychologii potocznej oznacza jednak nie tylko możliwość abstrahowania od istnienia rodzajowo identycznego *komponentu* systemu poznawczego, który realizowałby daną postawę propozycjonalną, ale także od istnienia GSS stanowiącego realizator tej postawy. Postuluję tu więc, że istnienie przekonań i pragnień jest niezależne od tego, czy dla dowolnej postawy propozycjonalnej będzie tak, że istnieje jeden (rodzajowo) GSS, który odpowiada przyczynowo za wszystkie manifestacje tej postawy. Analogicznie, zakładam, że prawdziwość czy poprawność osobowych *wyjaśnień* przyczynowych jest niezależna od tego, czy istnieje jakiś rodzajowo identyczny GSS, który wyjaśniałby przyczynowo wszystkie i tylko te działania, które – z punktu widzenia dostarczanego przez psychologię potoczną – stanowią manifestacje rodzajowo identycznej postawy propozycjonalnej.

Spróbujmy rozjaśnić powyższe stwierdzenia, nadając im bardziej konkretny kształt. Powiedzmy, że Jan dwa razy – odpowiednio w jakimś momencie t_1 oraz t_2 – w drodze do pracy zdecydował się wyjąć parasol z torby. Za każdym razem zrobił to pod wpływem (rodzajowo identycznego) pragnienia, aby nie zmoknąć[59]. W obu przypadkach działanie Jana zostało jednak poprzedzone i wywołane przez nomologicznie wystarczający warunek w postaci pewnego

[59] Rzecz jasna należy tu brać pod uwagę, że pragnienie to po części zawdzięcza swoją rolę przyczynową innym, towarzyszącym mu stanom mentalnym (trzeba założyć, że Jan widzi, że pada deszcz; sądzi, że pada deszcz i tak dalej).

GSS – odpowiednio GSS_1 (w t_1) i GSS_2 (w t_2). Przyjmijmy zarazem, że GSS_1 i GSS_2 nie są rodzajowo identyczne. To znaczy, że globalny stan, w jakim znajdowała się subosobowa architektura systemu poznawczego (Jana) w t_1, nie jest rodzajowo tożsamy ze stanem, w jakim znajdowała się ona w t_2. Mamy zatem cztery następujące horyzontalne wyjaśnienia przyczynowe:

(W1) W czasie t_1, działanie D Jana, polegające na wyjęciu parasola, zostało wywołane przez pragnienie P Jana, aby nie zmoknąć.
(W1′) W czasie t_1, D zostało wywołane przez GSS_1.
(W2) W czasie t_2, D zostało wywołane przez P.
(W2′) W czasie t_2, D zostało wywołane przez GSS_2.

Zgodnie z (W1) i (W2) działanie Jana miało rodzajowo identyczną przyczynę w t_1 i t_2. Jeśli za poprawne uznamy wyjaśnienia (W1′) i (W2′), teza taka musi zostać odrzucona i powinniśmy uznać, że wyjęcie parasola przez Jana każdego dnia miało rodzajowo różną przyczynę. Przywiązanie do mechanistycznej neutralności psychologii potocznej połączone z traktowaniem postaw propozycjonalnych jako przyczynowo efektywnych własności wyższego rzędu sprawia, że powinniśmy preferować (W1) kosztem (W1′) oraz (W2) kosztem (W2′). Jak można jednak uzasadnić tego rodzaju preferencję? Czy nie powinniśmy uznać po prostu, że to (W1′) i (W2′) wyjaśniają działania Jana?

Odpowiadając na wszystkie te wątpliwości, muszę przede wszystkim poczynić ważną klaryfikację, jeśli chodzi o bronione tu stanowisko. Należy wyraźnie zaznaczyć, jakiej tezy nie pragnę utrzymywać. Moją intencją nie jest negowanie oczywistego faktu, że GSS mogą wpływać przyczynowo na działania, gdzie przez „wpływ przyczynowy" rozumie się *bycie warunkiem nomologicznie wystarczającym* dla podjęcia przez kogoś określonego działania (por. Woodward 2008b). Wymienione wyżej zdania (W1′) i (W2′) są prawdziwe, jeśli interpretować je jako wskazujące takie nomologicznie wystarczające warunki wyjęcia przez Jana parasola w t_1 i t_2. Co więcej, kiedy twierdzę, że stany intencjonalne jako własności wyższego rzędu są przyczynowo efektywne, nie mam na myśli, że wpływają one na działania

dzięki temu, iż stanowią jakieś metafizycznie odrębne byty, aktywne „ponad" wewnętrzną, subosobową architekturą systemu poznawczego. Każde działanie dowolnego mechanizmu czy systemu (zakładając, że nie jest on marsjańską marionetką) jest determinowane przyczynowo zorganizowanym działaniem jego komponentów, albo też zorganizowanym działaniem komponentów oraz wystąpieniem pewnych okoliczności zewnętrznych. Atak serca – własność systemowa tego organu jako mechanizmu – nie może spowodować czyjejś śmierci inaczej, niż tylko przez bycie określonym zaburzeniem zorganizowanego działania komponentów serca. Przyczynowość własności systemowych jest zawsze w sposób nieunikniony „mechanistycznie zapośredniczona", w tym sensie, że to określony stan wewnętrznej, mechanistycznej organizacji (działających komponentów) stanowi zawsze warunek nomologicznie wystarczający do tego, by mechanizm czy system jako całość zachował się w określony sposób (por. Craver, Bechtel 2006)[60]. Analogicznie, przekonania i pragnienia mogą być warunkami nomologicznie wystarczającymi dla podjęcia określonych działań tylko „poprzez" GSS, a nie „obok" czy „ponad" nimi. Można też sformułować tę tezę w następujący sposób. Przypisując efektywność przyczynową przekonaniom i pragnie-

[60] Choć wyrażona tu teza jest zbieżna z koncepcją sformułowaną przez Cravera i Bechtela w przytaczanym artykule, zachodzi również rozbieżność, którą należy podkreślić. Wedle tych autorów przyczynowość własności systemowych jest możliwa dzięki temu, że są one konstytuowane przez określone stany wewnętrzne danego mechanizmu/systemu. Jeśli przez „konstytucję" rozumie się „konstytutywną relewantność" w sensie Cravera (por. sekcja 2.1.1), to teza ta nie jest problematyczna. Jeśli jednak Craver i Bechtel mają na myśli konstytucję w sensie zbliżonym do tego, jak pojmuje ją McDowell – na przykład jeśli twierdzą, iż stany mechanizmów są (rodzajowo/typicznie) identyczne z własnościami wyższego rzędu, albo że własności te są indywiduowane przez ich subosobowe odpowiedniki – to sąd taki musi być tu odrzucony. W pracy tej zgadzam się z tezą Cravera i Bechtela, że przyczynowość własności systemowych jest mechanistycznie zapośredniczona, lecz jednocześnie odrzucam twierdzenie, że relacja między stanami wewnętrznymi mechanizmu a własnościami wyższego rzędu (w szczególności postawami propozycjonalnymi) jest relacją konstytucji w sensie McDowella. Postawy propozycjonalne *wpływają* na zjawiska w sposób mechanistycznie zapośredniczony, jednak są *konstytuowane* przez dyspozycje systemu poznawczego jako całości, a nie przez jego stany wewnętrzne.

niom jako własnościom dyspozycyjnym, nie chcę tu negować faktu, że w pewnym istotnym sensie to *baza kategorialna* owych własności odpowiada za podejmowanie działań stanowiących ich manifestacje. Przekonania i pragnienia nie „wykraczają" przyczynowo poza ich bazę kategorialną – nie mogą one kształtować działania systemu poznawczego niezależnie od tej bazy albo inaczej niż jedynie *przez* nią (por. Schwitzgebel 2002).

Powyższe stwierdzenia mogą wydawać się znaczącym osłabieniem, a być może nawet kompletną rezygnacją z idei, że postawy propozycjonalne jako dyspozycyjne własności wyższego rzędu są efektywne przyczynowo. Zakładając, że przekonania i pragnienia nie są identyczne (rodzajowo) z określonymi GSS (bazą kategorialną) – z powyższych ustaleń w sposób nieunikniony wynika, iż to GSS (baza kategorialna), a nie same osobowe stany mentalne stanowią rzeczywisty *locus* umysłowych mocy przyczynowych. Uwaga ta byłaby jednak poprawna tylko, gdybyśmy (1) bycie przyczyną utożsamili z byciem warunkiem nomologicznie wystarczającym, natomiast (2) wyjaśnienie przyczynowe (horyzontalne) utożsamili ze wskazaniem takiego nomologicznie wystarczającego warunku (Woodward 2008b; por.: Baker 1995: 93–150; Craver 2007: 217–227). W takim wypadku postawy propozycjonalne rzeczywiście okazywałyby się (w jakimś sensie) epifenomenami. Jednakże tego rodzaju koncepcja przyczynowości i wyjaśniania przyczynowego – będąca w istocie „spadkiem" po nomologiczno-dedukcyjnym modelu wyjaśniania – natrafia na mocne zarzuty (por. szersze omówienie tego zagadnienia w: Grobler 2006: 105–109; Woodward 2008b). Generuje ona szereg niepożądanych konsekwencji i paradoksów, nie oddając zarazem sprawiedliwości temu, jakie funkcje rzeczywiście spełnia pojęcie przyczynowości, tak w nauce, jak w poznaniu potocznym (Woodward 2003, 2008b). Być może pytając o przyczynową efektywność przekonań i pragnień, powinniśmy zrewidować to, co dokładnie rozumiemy przez „przyczynowość" i „wyjaśnianie przyczynowe" (Baker 1995: 93–150; Craver 2007: 217–227; Woodward 2008b).

Chcę tu podążyć za sugestią niektórych autorów, że pytanie o przyczynowość mentalną rozumianą w kategoriach nomologicznie wystarczających warunków działań powinno zostać zastąpio-

ne pytaniem o przyczynowość mentalną rozumianą jako *relewantność przyczynowa* (*causal relevance*) osobowych stanów mentalnych (por.: Baker 1995: 93–150; Craver 2007: 217–227; Woodward 2008b; Raatikainen 2010, 2013; Shapiro 2011). Rzecz jasna takie przesunięcie akcentów naturalnie rodzi pytanie o naturę relewantności przyczynowej, czyli o to, co znaczy, że pewien *X* jest relewantny przyczynowo dla *Y*. Otóż w toku dalszych rozważań za punkt oparcia obierzemy konkretną, bardzo wpływową teorię relewantności przyczynowej (będącą zarazem teorią wyjaśniania przyczynowego). Teorię, o której mowa, stanowi *interwencjonistyczna* koncepcja przyczynowości Jamesa Woodwarda (2003, 2008a; 2008b; por. także Craver 2007: 63–106, 217–227, gdzie autor uzgadnia interwencjonizm Woodwarda z mechanistycznym ujęciem wyjaśniania w neuronauce). Jak się wydaje, wspomniana teoria pozwala w jasny i dobrze ugruntowany sposób zrehabilitować tezę o przyczynowej efektywności postaw propozycjonalnych jako własności systemowych. Po pierwsze, w kontekście tej koncepcji możemy zasadnie przypisywać role przyczynowe postawom propozycjonalnym (pojmowanym jako własności dyspozycyjne systemów poznawczych) bez narażania się na argument z wyłączenia przyczynowego. Po drugie, na gruncie tej teorii można zasadnie orzec, że wyjaśnienia odwołujące się do osobowych stanów intencjonalnych mogą być nie tylko w zupełności poprawne (prawdziwe), ale mają także znaczącą przewagę nad wyjaśnieniami odwołującymi się do GSS (czy szerzej – do faktów z poziomu osobowego).

Omówienie teorii Woodwarda w sposób oddający sprawiedliwość wszystkim jej detalom i niuansom jest tu niemożliwe. Zamiast tego skupię się tylko na zasadniczych tezach tego autora. Jego koncepcja wspiera się na twierdzeniu, że zależności przyczynowe to związki, które mogą być potencjalnie wykorzystywane w celach związanych z manipulacją i kontrolą. *X* jest przyczynowo relewantne dla *Y* wtedy, gdy *X* „czyni różnicę" dla *Y*, to znaczy gdy manipulacje *X* w odpowiedni sposób wiążą się ze zmianami *Y*. Precyzyjniej:

> **(M)** *X* wpływa przyczynowo na *Y* (jest przyczynowo relewantne dla niego), jeśli istnieją takie okoliczności tła *O*, że gdyby doszło

> do jakiejś (pojedynczej) interwencji w wartość *X* w *O*, to *Y* uległoby zmianie (za: Woodward 2008b: 222).

(M) wymaga przynajmniej kilku rozjaśnień (por. Woodward 2003; 2008b). Po pierwsze, koncepcja Woodwarda uzależnia zachodzenie relacji przyczynowej między *X* a *Y* od wyniku potencjalnej *interwencji* w *X*. W takim ujęciu kiedy stwierdzamy, że *X* jest przyczynowo relewantne dla *Y*, mamy na myśli w istocie, że odpowiedni eksperyment pokazałby, iż manipulacje *X* są w określony sposób powiązane ze zmianami *Y*. Jak każdy dobrze zaprojektowany eksperyment, powinien on wykluczać możliwość, że ewentualna korelacja między *X* a *Y* jest wynikiem działania jakiejś zmiennej kontrolnej. Po drugie, Woodward nie twierdzi, że przeprowadzenie eksperymentu to jedyna strategia pozwalająca w praktyce odkryć związek przyczynowy. Jednak związek ten *rzeczywiście zachodzi*, o ile przeprowadzenie takiego eksperymentu dałoby określone wyniki. Po trzecie, interwencja, o której mowa w (M), nie musi być technologicznie czy praktycznie wykonalna dla istot ludzkich. Na przykład twierdzenie, że dinozaury wyginęły ze względu na uderzenie w Ziemię meteorytu, należy rozumieć w następujący sposób: gdybyśmy mogli dokonać interwencji zapobiegającej uderzeniu tego meteorytu, dinozaury nie wyginęłyby (w każdym razie nie w tym konkretnym momencie historii życia na Ziemi). Twierdzenia dotyczące zależności przyczynowych są sensowne dopóty, dopóki (i) mamy względną jasność co do tego, na czym polegałaby potencjalna interwencja w *X* oraz (ii) dysponujemy jakimś niezależnym sposobem ustalenia, co by się stało, gdyby doszło do takiej interwencji. Po czwarte, teoria Woodwarda ujmuje zależności przyczynowe jako zachodzące między zmiennymi, to jest własnościami czy wielkościami, które mogą przyjmować różne wartości. Teoria interwencjonistyczna dopuszcza również zachodzenie związków przyczynowych między zdarzeniami, które to przecież klasycznie były uznawane za agumenty relacji przyczynowych. Powinniśmy je zdaniem Woodwarda rozumieć jako zmienne, które mogą przyjmować dwie wartości, odpowiadające zajściu oraz niezajściu zdarzenia. Po piąte, na gruncie omawianej teorii *explicite* przyjmuje się, że zależności przyczynowe zachodzą tylko i wy-

łącznie w pewnych okolicznościach tła (na przykład w atmosferze bogatej w tlen albo przy braku oddziaływań grawitacyjnych). Przyjmuje się jednak, że naukowo wartościowe zależności przyczynowe powinny być możliwie stabilne, to znaczy „niewrażliwe" na modyfikowanie okoliczności tła. Po szóste, na gruncie koncepcji manipulacyjnej zależności przyczynowe są określane przez to, w jaki sposób zmiany, którym podlega zmienna *X*, skutkują zmianami, jakim podlega zmienna *Y*. Nie wystarczy stwierdzić, że interwencje w *X* „jakoś" zmieniają wartości *Y*. Twierdzenia o przyczynowej relewantności zmiennej *X* dla *Y* mogą być oceniane jako prawdziwe lub fałszywe tylko wtedy, gdy *explicite* jest w nich wyrażone, które możliwe wartości zmiennej *X* powinny odpowiadać którym możliwym wartościom zmiennej *Y*.

Skupmy się teraz na konsekwencjach, jakie interwencjonizm niesie dla pytania o przyczynową efektywność postaw propozycjonalnych jako własności wyższego rzędu. Czy zależności przyczynowe, w których rolę miałyby odgrywać przekonania i pragnienia, spełniają warunek wymieniony w (M)? Czy interweniowanie w osobowe stany intencjonalne pozwala nam na manipulowanie jakimiś innymi zmiennymi? Wydaje się, że tak (por. Woodward 2008b). Każdego miesiąca psychologowie wykonują na świecie setki eksperymentów, w których jako zmienna niezależna występują postawy propozycjonalne osób badanych. Przykładami eksperymentów tego rodzaju są wypełnione chociażby podręczniki psychologii społecznej (por. na przykład Aronson, Wilson, Akert 1997). Co więcej, codzienne ludzkie interakcje zawierają miliony sytuacji, w których ludzie wpływają na cudze lub własne (na przykład w aktach autooszustwa; por. Hippel, Trivers 2011) przekonania, pragnienia, oczekiwania czy intencje w celu uzyskania jakiegoś efektu. Zarówno kontrolowane eksperymenty, jak i stosowane na co dzień manipulacje wykorzystują często zbliżone rodzaje środków, do których zaliczają się chociażby (w przypadku przekonań): werbalne informowanie, że *p*, przedstawianie argumentów czy racji za tym, że *p*, zmiana okoliczności percepcyjnych na takie, które (pozornie) świadczą, że *p*. Wydaje się, iż zarówno w nauce, jak i w życiu codziennym takie interwencje w postawy propozycjonalne pozwalają na otrzymywanie określonych

skutków, począwszy od zmiany czyjejś decyzji w sprawie filmu, jaki obejrzy w kinie, a skończywszy na upadkach imperiów.

Na gruncie interwencjonistycznej koncepcji przyczynowości nie ma zatem *prima facie* żadnych przeciwwskazań, by uznać, że postawy propozycjonalne są relewantne przyczynowo (por.: Woodward 2008b; Raatikainen 2010, 2013; Shapiro 2011). Aby uzasadnić tezę, że postawy propozycjonalne odgrywają role przyczynowe, wystarczy pokazać, iż odpowiednio manipulując tymi postawami, możemy wpływać na inne zjawiska. W szczególności możemy ustalać zachodzenie takich związków przyczynowych, całkowicie abstrahując od subosobowej architektury systemów poznawczych, w tym od GSS. Na przykład możemy racjonalnie i prawdziwie utrzymywać, że pragnienie Jana, aby nie zmoknąć, spowodowało wyjęcie przez niego parasola, pozostając zarazem kompletnie neutralnymi względem faktów zachodzących na poziomie subosobowym. Aby stwierdzić istnienie tej zależności przyczynowej, wystarczy wykazać, że interwencja w pragnienie Jana (sprawienie w jakiś sposób, by pragnął zmoknąć) pozwoliłaby nam zmienić jego decyzję w kwestii wyjęcia parasola. Właśnie z tego powodu na gruncie teorii interwencjonistycznej argument z wykluczenia przyczynowego nie może być wystosowany przeciwko przekonaniom i pragnieniom jako własnościom systemowym. Dla bycia przez *X* przyczynowo relewantnym dla *Y* wystarczy to, że manipulowanie *X* „czyni różnicę" dla *Y*. Kwestia natury realizatorów czy bazy kategorialnej *X* na niższym poziomie organizacji nie wchodzi tu po prostu w obręb rozważań. Związki przyczynowe z wyższych poziomów są stabilne nawet pomimo (ewentualnej) heterogeniczności realizatorów z niższego poziomu, stanowiąc tym samym „związki zależnościowe niezależne od realizacji" (*realization independent dependency relationships*; Woodward 2008b).

Spójrzmy teraz na dwa poniższe wyjaśnienia:

(W1) W czasie t_1 działanie D Jana, polegające na wyjęciu parasola, zostało wywołane przez pragnienie P Jana, aby nie zmoknąć.
(W1′) W czasie t_1 D zostało wywołane przez GSS_1.

Kiedy poruszamy się na gruncie koncepcji interwencjonistycznej oraz odróżniamy relewantność przyczynową od bycia warunkiem nomologicznie wystarczającym, akceptacja (W1′) nie wyklucza akceptacji (W1). Prawdziwość (W1) wymaga wyłącznie tego, aby określone interwencje w pragnienie Jana pozwalały manipulować jego działaniem (wyjęciem parasola). Prawdziwości (W1) nic nie „zagraża", o ile tylko dysponujemy ogólną ideą, na czym polegałoby manipulowanie pragnieniem Jana, oraz jesteśmy w stanie określić, jak tego rodzaju manipulacje wpłynęłyby na to, czy Jan wyjął parasol.

Powyższe konstatacje nie rozwiązują jeszcze jednak kolejnego wymienionego wcześniej problemu, który powstaje, kiedy zestawimy cztery następujące horyzontalne wyjaśnienia przyczynowe:

(W1) W czasie t_1 działanie D Jana, polegające na wyjęciu parasola, zostało wywołane przez pragnienie P Jana, aby nie zmoknąć.
(W1′) W czasie t_1 D zostało wywołane przez GSS_1.
(W2) W czasie t_2 D zostało wywołane przez P.
(W2′) W czasie t_2 D zostało wywołane przez GSS_2.

Jeśli przekonania i pragnienia rzeczywiście mają stanowić eksplananse w poprawnych i wartościowych horyzontalnych wyjaśnieniach przyczynowych, to powinny istnieć podstawy, by preferować wyjaśnienia intencjonalne w rodzaju (W1) i (W2) kosztem tych w rodzaju (W1′) lub (W2′). Dlaczego powinniśmy preferować bardziej „gruboziarniste" wyjaśnienia psychologii potocznej kosztem bardziej „drobnoziarnistych" wyjaśnień odwołujących się do GSS? Koncepcja interwencjonistyczna dostarcza odpowiedzi na to pytanie. Wartościowe wyjaśnienia przyczynowe powinny zawsze nieść pewne konsekwencje dotyczące tego, „co by było, gdyby" przyczyna znalazła się w stanie innym niż ten, który rzeczywiście doprowadził do określonego skutku (Woodward 2003, 2008b, 2010; por. także omówienie zagadnienia klas kontrastu w wyjaśnieniach naukowych przedstawione w: Grobler 2006). Wyjaśnienie mówiące, że przyjęcie wartości x_i przez zmienną X sprawiło, iż wartość zmiennej Y wyniosła y_i powinno ujmować wzorzec zależności między X a Y o takiej „ziarnistości", by można było orzec, że *gdyby* X nie przyjęło

wartości x_i (lecz jakąś inną wartość x_j), to Y nie przyjęłoby wartości y_i (lecz jakąś inną wartość y_j). Możemy powiedzieć, że przyczyna powinna być *proporcjonalna* do skutku (Woodward 2008b, 2010; por. Yablo 1992). A zatem powinien być spełniony następujący warunek:

> **(P)** Istnieje wzorzec systematycznych kontrfaktycznych zależności (rozumianych w duchu interwencjonizmu) między różnymi możliwymi stanami przyczyny a różnymi możliwymi stanami skutku, który to wzorzec zależności w przybliżeniu odpowiada następującemu ideałowi: zależność (jej charakterystyka) powinna być taka, że (a) *explicite* lub *implicite* zawiera poprawne informacje na temat warunków, w których skutek znajdzie się w określonych alternatywnych stanach *oraz* (b) zawiera *tylko* te informacje – to znaczy przyczyna nie jest scharakteryzowana w taki sposób, że jej alternatywne stany *nie* będą związane ze zmianami w stanach skutku (za: Woodward 2010: 298).

Powyższą myśl można zilustrować prostym przykładem, zaczerpniętym przez Woodwarda (2008b, 2010) z pracy Stephena Yablo (1992). Załóżmy, że pewien gołąb został uwarunkowany w taki sposób, aby dziobać za każdym razem, gdy jest mu przedstawiany bodziec koloru czerwonego. Wyobraźmy sobie, że w danej sytuacji gołąb zaczyna dziobać po tym, jak zaprezentowano mu bodziec w jakimś określonym odcieniu koloru czerwonego, na przykład szkarłatnym. Rozważmy teraz dwa możliwe wyjaśnienia tego, dlaczego gołąb zaczął dziobać w tej sytuacji:

> **(G)** Prezentacja czerwonego bodźca sprawiła, że gołąb zaczął dziobać.
> **(G′)** Prezentacja szkarłatnego bodźca sprawiła, że gołąb zaczął dziobać.

Według Woodwarda (jak również Yablo) wyjaśnienie (G) powinno być preferowane kosztem (G′). Dlaczego? Otóż (G′) jest zbyt szczegółowe czy drobnoziarniste, przez co nie spełnia ono warunku (b) zawartego w (P). (G′) zawiera szczegóły nieistotne dla wzorca za-

leżności przyczynowej, która wyjaśnia zachowanie gołębia. Nie jest tak, że gołąb nie zacząłby dziobać, gdyby prezentowany mu bodziec nie był szkarłatny. Dowolny odcień czerwonego sprawiłby, że gołąb zacznie dziobać. To nie fakt, iż bodziec jest szkarłatny (a nie – nie-szkarłatny), lecz to, że jest czerwony (a nie – nie-czerwony) stanowi czynnik przyczynowo relewantny dla zachowania gołębia. Na tym właśnie polega wyższość wyjaśnienia (G): jest to wyjaśnienie na tyle szczegółowe, aby „wydobyć" ważną zależność ze świata, ale nie jest tak szczegółowe, by odwoływać się do czynników, które nie są relewantne dla określonego skutku. Skupiając się na nieistotnych szczegółach, (G′) *de facto* ignoruje ważną regularność czy związek, który występuje na wyższym poziomie „ziarnistości" kolorów (por. Woodward 2003, 2008b, 2010). Przyczyna wymieniona w (G) będzie więc proporcjonalna do skutku, a ta wymieniona w (G′) – nie.

Jak sądzę, na podobnej zasadzie można bronić wyższości wyjaśnień w rodzaju (W1) i (W2) nad wyjaśnieniami w rodzaju (W1′) i (W2′)[61]. Wyjaśnienia wyjęcia parasola przez Jana za pomocą GSS są *zbyt* drobnoziarniste i szczegółowe (por.: Baker 1995: 93–120; Woodward 2008b; Raatikainen 2010, 2013). Przypominają one pod tym względem (G′). Za działanie Jana mogą odpowiadać (w sensie:

[61] Aby upodobnić wymieniony wyżej przykład z gołębiem do przykładu dotyczącego przyczynowości mentalnej, można sobie wyobrazić, że w innym momencie zwierzę zaczęło dziobać po pokazaniu mu bodźca w cynobrowym odcieniu czerwonego. Moglibyśmy wtedy rozważyć nie dwa, a cztery następujące wyjaśnienia:

> **(G1)** W czasie *t1* prezentacja czerwonego bodźca sprawiła, że gołąb zaczął dziobać.
> **(G1′)** W czasie *t1* prezentacja szkarłatnego bodźca sprawiła, że gołąb zaczął dziobać.
> **(G2)** W czasie *t2* prezentacja czerwonego bodźca sprawiła, że gołąb zaczął dziobać.
> **(G2′)** W czasie *t2* prezentacja cynobrowego bodźca sprawiła, że gołąb zaczął dziobać.

Ze względów wymienionych w tekście głównym, (G1) i (G2) powinny być preferowane kosztem (G1′) i (G2′). Na analogicznej zasadzie (W1) i (W2) są lepszymi wyjaśnieniami od (W1′) i (W2′).

być nomologicznie wystarczające dla jego wykonania) bardzo heterogeniczne okoliczności wewnętrzne[62]. Nie jest tak, lub nie musi tak być, że Jan *nie* podejmie działania D, o ile nie znajdzie się on w GSS_1, ponieważ do podjęcia tego działania mógłby doprowadzić GSS_2, a zapewne także jakiś GSS_3, GSS_4 i tak dalej. Przypomnijmy więc sobie, że *Y* to zmienna, która przyjmuje dwie wartości: „Jan wyjął parasol; Jan nie wyjął parasola". Załóżmy też, że chcemy *Y* wyjaśnić przyczynowo (horyzontalnie) za pomocą subosobowej zmiennej *X*, której możliwe wartości to „GSS_1, GSS_2, GSS_3, ..., GSS_n". W zgodzie z punktem (b) w (P) – wyjaśnienie przyczynowe *X* przez *Y* wymagałoby pokazania, że manipulowanie *X* będzie wywoływać *proporcjonalne* zmiany w *Y*. Tymczasem w omawianym przypadku zmienna *X* zawiera wiele wartości, które nie „czynią różnicy" dla *Y*. Istnieje więc wiele GSS, których wystąpienie będzie mieć ten sam skutek. Wartości zmiennej *Y* „Jan wyjął parasol" odpowiada wiele różnych możliwych wartości zmiennej *X*. Tym samym w wyjaśnieniach (W1′) i (W2′) przyczyna nie jest proporcjonalna do skutku.

Wyjaśnianie wyjęcia parasola przez Jana za pomocą GSS nie spełnia więc ważnego wymogu nakładanego przez koncepcję interwencjonistyczną na horyzontalne wyjaśnienia przyczynowe. Wydaje się, że wzorce zależności poprawnie wyjaśniające działanie Jana pojawiają się dopiero na *wyższym poziomie*. Jakim konkretnie? Postuluję, że jest to poziom psychologii potocznej, a zatem poziom organizacji, na którym mówimy o systemie poznawczym jako całości zaangażowanej w interakcje z zamieszkiwanym przezeń światem. To właśnie tu odnajdujemy bowiem, jak sądzę, regularną zależność, dla której punkt (b) z (P) jest spełniony (por. Woodward 2008b). Dopie-

[62] Powtórzę tu: nie wykluczam *a priori*, że taksonomia GSS (globalnych, mechanistycznych realizatorów postaw propozycjonalnych) będzie odpowiadać taksonomii psychologii potocznej – to znaczy, że każda pojedyncza (rodzajowo) postawa propozycjonalna będzie miała pojedynczy (rodzajowo) odpowiednik w postaci GSS. Jednak twierdzę, że przyczynowa efektywność postaw propozycjonalnych pozostaje kompletnie niezależna od tego, czy tak rzeczywiście będzie. W szczególności jest ona (a w każdym razie powinna być, jeśli prowadzone tu rozważania idą dobrym torem) niezagrożona nawet wtedy, gdy taka odpowiedniość międzypoziomowa nie zachodzi.

ro na tym poziomie „lokuje się" pragnienie Jana, by nie zmoknąć. To właśnie odwołanie się do pragnienia jako zmiennej niezależnej spełnia warunki nałożone na wyjaśnianie przyczynowe przez teorię Woodwarda. Aby poprawnie wyjaśniać *Y* przyjmujące wartości: „Jan wyjął parasol; Jan nie wyjął parasola", musimy odwołać się do zmiennej *X*, która przyjmuje wartości: „Jan pragnie, aby nie zmoknąć; Jan nie pragnie, aby nie zmoknąć"[63]. W takiej sytuacji przyczyna okazuje się proporcjonalna do skutku. Przykład ten pokazuje, że postawy propozycjonalne jako własności systemowe mogą stanowić eksplananse w poprawnych, wartościowych horyzontalnych wyjaśnieniach przyczynowych.

[63] W tym przykładzie objawia się także warta odnotowania słabość przyczynowego wyjaśniania zjawisk za pomocą postaw propozycjonalnych. Jak zostało wcześniej zaznaczone, przyczynowe wyjaśnienia powinny być według Woodwarda (2010) stabilne, to znaczy – odwoływać się do zależności, które są możliwie niezależne od zmian okoliczności tła. Tymczasem role przyczynowe przekonań i pragnień wydają się ściśle zależne od okoliczności tła wytwarzanych przez cały szereg innych osobowych stanów mentalnych danej osoby. Na przykład jeśli Jan jest przekonany, iż jego parasol jest dziurawy, to może on nie wyjąć parasola nawet wtedy, gdy pragnie on, by nie zmoknąć. Przyczynowość mentalna wydaje się w nieunikniony sposób „czuła" na kontekst wyznaczany przez inne stany intencjonalne. Tym samym związki przyczynowe, w jakie wchodzą postawy propozycjonalne, należy chyba uznać za mało stabilne.

Zakończenie

W rozdziale 1 postawiłem diagnozę dotyczącą zamieszania pojęciowego, jakie towarzyszy współczesnej debacie na temat istnienia reprezentacji mentalnych oraz roli eksplanacyjnej, jaką mają one do spełnienia w naukach kognitywnych. Zwróciłem uwagę na fakt, że poszczególni uczestnicy tej debaty – zarówno reprezentacjoniści, jak i antyreprezentacjoniści – różnie rozumieją termin „reprezentacja", a często posługują się nim w sposób na tyle liberalny i teoretycznie niezobowiązujący, że przestaje on desygnować jakąkolwiek eksplanacyjnie użyteczną kategorię. Dodatkowo kwestia natury wyjaśnień reprezentacyjnych w kognitywistyce nie została w wystarczającym stopniu podjęta przez filozofów umysłu, którzy skupili się na – przynajmniej częściowo osobnym i niezależnym – zagadnieniu naturalizacji intencjonalności. Sytuacji nie poprawia też fakt, iż wiele wpływowych filozoficznych prób naturalizacji intencjonalności wydawało się opierać na (bardziej lub mniej ukrytym) założeniu, jakoby rola eksplanacyjna, jaką pojęcie reprezentacji mentalnych odgrywa w manifestującym się obrazie świata (psychologii potocznej), nie różniła się w zasadzie od roli odgrywanej przez nie w obrazie naukowym (kognitywistyce).

Zasadniczym celem tej książki było wprowadzenie jasności i porządku w ten konceptualny chaos panujący na pograniczu kognitywistyki i filozofii. Pora podsumować zarówno prowadzony tu wywód, jak i płynące z niego konkluzje.

W pracy tej skoncentrowałem się na trzech powiązanych ze sobą problemach. Głównym zagadnieniem tu podejmowanym był metaprzedmiotowy problem statusu eksplanacyjnego reprezentacji w kognitywistyce:

> **Problem 1 (główny):** Na czym polega wyjaśnienie danego zjawiska za pomocą reprezentacji? Jak odróżnić pełnoprawnie

reprezentacyjne wyjaśnienie od takiego, które nie jest reprezentacyjne (lub jest reprezentacyjne jedynie pozornie)? Na jakiej podstawie możemy stwierdzić, że badany system rzeczywiście posługuje się reprezentacjami?

Podejmując ten problem, za punkt wyjścia obrałem założenie, że wyjaśnianie w kognitywistyce stanowi formę wyjaśniania mechanistycznego. Pytanie o naturę wyjaśniania reprezentacyjnego w naukach o poznaniu zostało uszczegółowione jako pytanie o naturę *mechanistycznego* wyjaśniania reprezentacyjnego. Przyjąłem, że mechanistyczne wyjaśnienie reprezentacyjne pewnego eksplanandum to wyjaśnienie go za pomocą mechanizmu reprezentacyjnego, mającego (co najmniej jeden) komponent odgrywający rolę funkcjonalną, polegającą na reprezentowaniu czegoś. W celu określenia, jakie warunki (funkcjonalne) powinien spełnić komponent mechanizmu, aby był reprezentacją (pełnił funkcję reprezentacji), wykorzystałem procedurę nazwaną w tej książce „metodą Ramseya" (nawiązując do: Ramsey 2007). Metoda ta pozwala w teoretycznie umotywowany sposób odróżnić struktury pełniące w systemie poznawczym rolę reprezentacji od takich, które takiej roli nie pełnią – nawet jeśli są rutynowo nazywane „reprezentacjami" przez kognitywistów. Podążając taką ścieżką, sformułowałem rozwiązanie metaprzedmiotowego problemu statusu eksplanacyjnego reprezentacji. Zgodnie z przedstawioną tu propozycją reprezentacje mentalne w eksplanacyjnie wartościowym dla kognitywistyki sensie to konsumowane modele wewnętrzne. Z kolei mechanizmy reprezentacyjne to mechanizmy wykorzystujące owe modele w swoim działaniu. Nośnikiem takiego modelu jest zawsze komponent, którego poprawne funkcjonowanie w szerszym mechanizmie systematycznie zależy od tego, czy między nim samym (czyli nośnikiem) a pewną zewnętrzną domeną – przedmiotem reprezentacji – zachodzi relacja podobieństwa strukturalnego. Model mentalny ma zawsze swojego konsumenta, to znaczy towarzyszy mu inny komponent mechanizmu, który przy realizowaniu własnej funkcji wykorzystuje podobieństwo zachodzące między nośnikiem a przedmiotem reprezentacji. Proponuję więc następujące rozwiązanie głównego problemu poruszanego w tej pracy:

> **Proponowane rozwiązanie problemu 1 (teza główna):** Reprezentacyjne wyjaśnienie danego zjawiska w kognitywistyce to wyjaśnienie tego zjawiska za pomocą mechanizmu wyposażonego w konsumowany model (MKM) pewnej domeny. System poznawczy jest systemem reprezentacyjnym w takim zakresie, w jakim jego aktywność opiera się na tego rodzaju mechanizmach.

Takie rozwiązanie problemu reprezentacji na poziomie metaprzedmiotowym wykorzystałem następnie na poziomie *przedmiotowym*. Przedmiotowy problem statusu eksplanacyjnego reprezentacji stanowił bowiem drugie zagadnienie poruszane w tej pracy:

> **Problem 2:** Czy kognitywiści potrzebują pojęcia reprezentacji do realizacji stawianych sobie celów eksplanacyjnych? Czy dobre (poprawne, najlepsze spośród dostępnych) wyjaśnienia odwołują się do wewnętrznych reprezentacji? Czy (lub w jakim zakresie) system poznawczy jest systemem reprezentacyjnym?

Zgodnie z proponowaną tu koncepcją system poznawczy jest systemem reprezentacyjnym w takim zakresie, w jakim jego działanie opiera się na MKM. W celu rozwiązania wymienionego wyżej problemu zająłem się zatem kwestią tego, czy MKM rzeczywiście wyjaśniają jakieś zjawiska poznawcze w kognitywistyce. Skupiłem się konkretnie na krytyce tezy Williama Ramseya (2007), iż współczesna, „nieklasyczna" kognitywistyka w zasadzie nie powołuje się na modele mentalne jako ważne narzędzie eksplanacyjne. Próbowałem pokazać, że taka ocena sytuacji jest błędna. Mechanizmy reprezentacyjne (czyli MKM) pełnią istotną funkcję eksplanacyjną w wielu obszarach współczesnej kognitywistyki: w modelowaniu koneksjonistycznym, neuronauce poznawczej i obliczeniowej, robotyce poznawczej oraz w ramach podejścia opartego na teorii sterowania. Rzeczywista sytuacja teoretyczna współczesnej kognitywistyki sprzyja zatem tezie, że system poznawczy wykorzystuje w swoim działaniu mechanizmy wyposażone w konsumowane modele. Zakładając, iż współczesne teorie i wyjaśnienia aktywności systemu poznawczego dają nam rzeczywisty wgląd w jego strukturę i dzia-

łanie, musimy stwierdzić, że system ten jest przynajmniej w jakimś zakresie reprezentacyjny. Pozostawiam tu jednak otwartym zagadnienie, w jakim zakresie jest on reprezentacyjny: dopuszczam możliwość, że niektóre ważne eksplananda kognitywistyki nie podlegają wyjaśnieniu za pomocą MKM. Innymi słowy – sądzę, że reprezentacjonizm jest stanowiskiem poprawnym „lokalnie" (w odniesieniu do konkretnych eksplanandów), ale niekoniecznie „globalnie" (istnieją bowiem zapewne eksplananda, które nie mają u swoich podstaw reprezentacji). Drugi wniosek płynący z prowadzonych tu rozważań jest zatem następujący:

> **Proponowane rozwiązanie problemu 2:** Mechanizmy wyposażone w konsumowane modele pełnią ważną funkcję w repozytorium eksplanacyjnym współczesnej kognitywistyki. Tym samym stan teoretyczny nauk o poznaniu każe sądzić, że aktywność systemu poznawczego opiera się (w jakimś zakresie) na MKM. Można więc powiedzieć, że system poznawczy jest w pewnym zakresie systemem reprezentacyjnym.

Trzecim celem tej książki było zbadanie zależności między (1) kwestią statusu eksplanacyjnego reprezentacji w kognitywistyce a (2) problemem naturalizacji intencjonalności:

> **Problem 3:** Czy powodzenie projektu naturalizacji intencjonalności zależy od faktów dotyczących eksplanacyjnej wartości reprezentacji dla kognitywistów? Czy powodzenie projektu naturalizacji zależy od tego, jakie własności intencjonalne i funkcjonalne przysługują reprezentacjom postulowanym przez kognitywistów (zakładając, że reprezentacje są w ogóle przez nich postulowane)? Czy psychologia potoczna wymaga naukowej „legitymizacji"?

Podejmując ten problem, za punkt oparcia raz jeszcze przyjąłem mechanistyczny model wyjaśniania. Wykorzystałem mechanicyzm, aby nadać interpretację klasycznemu rozróżnieniu na osobowy oraz subosobowy poziom wyjaśniania systemu poznawczego. Zinterpre-

towałem poziom osobowy – czyli poziom, na którym jest aplikowana aparatura pojęciowa psychologii potocznej – jako „ekologiczny", skupiający się na systemie poznawczym jako całości zaangażowanej w różnorakie interakcje z zamieszkiwanym środowiskiem. Natomiast poziom subosobowy zinterpretowałem jako taki, na którym są formułowane mechanistyczne wyjaśnienia zdolności poznawczych. Wyjaśnienia te odwołują się do niższych „warstw" hierarchicznej, mechanistycznej organizacji systemu poznawczego. Centralnym elementem bronionej tu propozycji była teza, że wyjaśnienia z poziomu osobowego są mechanistycznie neutralne. Wyjaśnienia działań za pomocą psychologii potocznej nie niosą ze sobą architekturalnych zobowiązań dotyczących wewnętrznej, mechanistycznej organizacji systemu poznawczego. Z kolei stanowiące przedmiot projektu naturalizacji osobowe stany intencjonalne – przekonania, pragnienia, intencje i tak dalej – są własnościami wyższego rzędu, egzemplifikowanymi przez systemy poznawcze jako całości, a nie przez wewnętrzne komponenty tych systemów. Kognitywistyka w swoich wyjaśnieniach odwołuje się do reprezentacji subosobowych (będących komponentami mechanizmów poznawczych), natomiast psychologia potoczna – do osobowych stanów intencjonalnych. Są to zupełnie różne rodzaje struktur czy stanów, spełniające różne funkcje eksplanacyjne. Takie postawienie sprawy prowadzi (między innymi) do wniosku, iż problem naturalizacji intencjonalności jest w znacznym stopniu autonomiczny względem kwestii użyteczności eksplanacyjnej (subosobowych) reprezentacji w kognitywistyce:

> **Proponowane rozwiązanie problemu 3:** Psychologia potoczna spełnia inne funkcje eksplanacyjne niż kognitywistyka. Przekonania, pragnienia i inne postawy propozycjonalne – stanowiące przedmiot zainteresowania projektu naturalizacji intencjonalności – nie odgrywają roli w subosobowych, mechanistycznych wyjaśnieniach kognitywistyki. Widnieją one w wyjaśnieniach skoncentrowanych na wyższym, osobowym poziomie organizacji systemu poznawczego. Osobowe stany intencjonalne (reprezentacje osobowe) to własności systemowe, które nie są „zlokalizowane" na poziomie subosobowym. To, czy reprezentacje

> subosobowe okażą się eksplanacyjnie użyteczne dla kognitywistów, nie decyduje o prawomocności bądź braku prawomocności wyjaśniania ludzkich działań za pomocą przekonań i pragnień. Powodzenie projektu naturalizacji intencjonalności nie zależy od sukcesu reprezentacjonizmu w kognitywistyce, a psychologia potoczna nie wymaga tego rodzaju naukowej „legitymizacji".

O wartości zarówno teorii naukowych, jak i filozoficznych często decyduje nie tylko to, czy pozwalają one definitywnie rozwiązać jakiś problem, lecz także to, czy otwierają nowe perspektywy badawcze i jaki nadają kierunek dalszym poszukiwaniom. Odpowiedzi udzielone na jakieś pytanie są czasem wartościowe także dlatego, że owocnie ukierunkowują dalszy proces stawiania pytań. Przed zamknięciem rozważań warto zapytać, jakie dalsze zagadnienia i perspektywy badawcze otwierają się przed nami – w odniesieniu do każdego z trzech problemów, które nas tu zajmowały – jeśli zaakceptujemy zaproponowane w tej pracy rozwiązania.

Zacznijmy od metaprzedmiotowego problemu statusu eksplanacyjnego reprezentacji. Na gruncie proponowanych tu rozwiązań powstają dwa dalsze zagadnienia, które zostały wcześniej zaledwie zaznaczone (sekcja 4.2.2). Pierwsza z tych kwestii dotyczy tego, czy MKM są jedynym rodzajem mechanizmów reprezentacyjnych. Prowadzone tu rozważania każą wyciągnąć wniosek, że spośród wszystkich struktur konceptualizowanych przez kognitywistów jako reprezentacje, tylko modele mentalne w sposób niekontrowersyjny spełniają Ramseyowski wymóg opisu zadań. Pod nieobecność argumentów przeciwko takiemu stanowisku, niejako „roboczo" zaakceptowałem tu twierdzenie o eksplanacyjnej „wyjątkowości" modeli mentalnych – w tym sensie, że uznałem takie modele za jedyny eksplanacyjnie prawomocny rodzaj wewnętrznych reprezentacji. Moim celem nie było jednak definitywne wykluczenie możliwości, iż także inne rodzaje struktur mogą z powodzeniem przejść test oparty na wymogu opisu zadań. To, czy tak rzeczywiście jest oraz jakie inne rodzaje reprezentacji są dostępne kognitywistom, pozostaje jednak otwarte. Temat ten wymaga bez wątpienia dalszej eksploracji.

Drugie zagadnienie wyłaniające się na poziomie metaprzedmiotowym to problem możliwych rozszerzeń i uzupełnień teorii MKM. Jak wspomniałem w rozdziale 4, MKM mogą posiadać pewne dodatkowe własności, które nie zostały wymienione w charakterystyce przedstawionej w sekcji 4.2.1. Jako przykład takiej dodatkowej własności podałem zdolność do wykrywania błędu reprezentacyjnego (por. także Gładziejewski, w druku). Możliwość rozszerzania bronionej w tej pracy teorii mechanizmów reprezentacyjnych jest istotna ze względu na pewną dwoistość dotyczącą samej idei modeli mentalnych. Otóż przedstawiona tu *explicite* charakterystyka tych modeli okazuje się stosunkowo minimalistyczna. Warunki nakładane na bycie MKM są na tyle szerokie, że może je spełniać bardzo wiele różnych mechanizmów. Modele w przyjmowanym tu znaczeniu mogą wyjaśniać cały szereg funkcji czy zdolności poznawczych, wliczając w to bardzo proste i filogenetycznie stare funkcje związane z kontrolą motoryczną. Ángel García i Paco Calvo (2010) zwracają uwagę, że z tak ogólnie rozumianych modeli korzystają nawet przedstawiciele świata roślin, którzy do przewidywania pozycji słońca o poranku wykorzystują wewnętrzne, biologiczne oscylatory dynamicznie odzwierciedlające przebieg rytmu dobowego[1]. Tak szerokie ujęcie modeli mentalnych może być rozczarowujące dla niektórych filozofów. Mogą oni oczekiwać, że prawdziwie interesujące filozoficznie modele mentalne to struktury związane z wyższymi, być może specyficznie ludzkimi zdolnościami umysłowymi, takimi jak prospekcja, pamięć autobiograficzna, myślenie kontrfaktyczne, poznanie moralne, przeprowadzanie rozumowań czy posługiwanie się językiem (por.

[1] Co bardzo istotne, oscylatory te działają w sposób mocno oderwany, to znaczy pełnią swoją funkcję, kiedy roślina nie interreaguje przyczynowo ze Słońcem. Wydaje się, że w świetle argumentacji Garcíi i Calvo należy uznać niektóre rośliny nie tylko za systemy (minimalnie) poznawcze, ale także systemy reprezentacyjne. Jest to z pewnością konstatacja niezwykle kontrintuicyjna i filozoficznie zaskakująca, jednak dobrze ugruntowana w faktach dotyczących działania przedstawicieli świata roślin. Trzeba jednak bardzo wyraźnie podkreślić: z faktu, że (1) rośliny posługują się wewnętrznymi modelami, nie wynika, iż (2) nie zachodzą teoretycznie istotne różnice między takimi modelami a modelami, z jakich korzystają biologiczne mózgi angażujące się w realizowanie wyższych (w tym specyficznie ludzkich) zdolności poznawczych.

Morgan 2014). Przykładem modelu mentalnego w takim filozoficznie „maksymalistycznym" znaczeniu jest postulowany przez Roberta Piłata (1999) funkcjonujący, osobisty model świata. W ujęciu tego autora system poznawczy wytwarza pewne modele „szczątkowe" (których przedmiotem są fragmenty świata) – integrowane w ogólny, osobisty model świata danej osoby. Jak pisze Piłat, łączenie modeli „[...] w osobisty model świata następuje w ciągu rozwoju danej osoby, tworząc stopniowo indywidualny styl poznawczy i osobisty świat wartości" (1999: 12). Nie ulega wątpliwości, że takie rozumienie modelu mentalnego odbiega pod istotnymi względami od ujęcia, które za modele każe uznawać także relatywnie „trywialne" filozoficznie struktury, uczestniczące w kontroli okulomotorycznej czy odpowiadające za dostosowanie się roślin do pozycji Słońca na niebie. Jednakże właśnie z tego powodu tak ważne pozostaje podkreślenie, że broniona tu teoria formułuje tylko zupełnie *elementarne* czy *minimalne* warunki bycia MKM. Sądzę, że koncepcja ta poakzuje, co czyni pewną wewnętrzną strukturę *modelem* (a przez to także *reprezentacją*) *per se*, jednak nie zaprzeczam, że mogą istnieć *różne rodzaje* modeli mentalnych. Zarówno proste, minimalnie poznawcze funkcje związane z kontrolą motoryczną, jak i wyższe, specyficznie ludzkie funkcje wykorzystują modele mentalne w tym samym zasadniczym znaczeniu „modelu" jako wewnętrznej, konsumowanej S-reprezentacji[2]. Nie zmienia to w najmniejszym stopniu dość oczywistego faktu, że modele uczestniczące w wyższych, typowo ludzkich zdolnościach poznawczych mają jakieś cechy dodatkowe, które nie zostały tu wymienione. Inaczej mówiąc, modele odpowiadające za, dla przykładu, myślenie kontrfaktyczne będą bez wątpienia posiadać pewne specyficzne własności, które nie przysługują (dajmy na to) modelom, z jakich korzysta system okulomotoryczny. Jakie to jednak własności? Czym różnią się modele mentalne uczestniczące w realizowaniu wyższych funkcji poznawczych od prostszych rodzajów modeli? Czy różnice te są związane ze złożonością nośników

[2] Warto na marginesie zauważyć, że w rozdziale 4 (podrozdział 4.3) jako przykłady MKM wymieniłem nie tylko mechanizmy odpowiadające za proste funkcje poznawcze, ale także chociażby mechanizm postulowany przez Barsalou do wyjaśnienia funkcji pojęciowych.

i przedmiotów reprezentacji? Czy dotyczą natury konsumentów? A może wynikają one z istnienia jakichś innych, dodatkowych własności, jakie przysługiwałyby pewnym rodzajom MKM, a nie innym? Są to właśnie pytania, które wymagają dalszych badań.

Jeśli zaś chodzi o przedmiotowy problem użyteczności eksplanacyjnej reprezentacji, dalsze badania powinny bez wątpienia skupić się na rzeczywistej roli MKM w funkcjonowaniu systemów poznawczych. Przedstawiony w tej pracy zestaw przykładów zastosowania MKM w nieklasycznej kognitywistyce stanowi zaledwie szkic rozległego tematu. Istnieje tu cała grupa otwartych problemów empirycznych wymagających systematycznego podjęcia. Które z MKM postulowanych współcześnie przez kognitywistów przetrwają próbę czasu i staną się częścią kompletnej teorii działania systemu poznawczego? Czy inicjowane współcześnie projekty zmierzające do stworzenia szczegółowych map strukturalnej i funkcjonalnej organizacji mózgu oraz obliczeniowych symulacji jego aktywności przyniosą rezultaty sprzyjające tezie o istnieniu MKM? Gdzie w mózgu są zlokalizowane MKM? Jak starą filogenetycznie strategią jest wykorzystywanie wewnętrznych modeli do realizowania funkcji poznawczych? Czy odwołujące się do MKM teorie wymienione w rozdziale 4 (podrozdział 4.3) są niezależne od siebie? A może przynajmniej niektóre z nich – na przykład teoria emulacji oraz wspomniana tu pobieżnie teoria kodowania predykcyjnego – opisują te same mechanizmy, tylko za pomocą innego języka teoretycznego? Czy wyjaśnienia odwołujące się do MKM mają zastosowanie lokalne (zrelatywizowane do poszczególnych eksplanandów), czy też jest możliwa unifikacja teoretyczna kognitywistyki pokazująca, że modele mentalne przenikają wszelkie formy aktywności systemu poznawczego? Wszystkie te zagadnienia wymagają skrupulatnych badań – nie tyle filozoficznych czy konceptualnych, co raczej empirycznych. Z prowadzonych tu rozważań wynika jednak, że od tego, jak odpowiemy na wymienione pytania, zależą losy reprezentacjonizmu w kognitywistyce.

Wreszcie jeśli chodzi o ostatni poruszany w tej pracy problem – zagadnienie relacji między rolą eksplanacyjną (subosobowych) reprezentacji w kognitywistyce a projektem naturalizacji intencjonalno-

ści – to należałoby rozwinąć i uzupełnić poszczególne rozwiązania zaproponowane w rozdziale 5. Przedstawiona tam (w podrozdziale 5.3) teoria postaw propozycjonalnych jako przyczynowo efektywnych własności dyspozycyjnych wymaga dopracowania. Powinien zostać rozwinięty chociażby wątek dotyczący nierelacyjnej, pomiarowej koncepcji treści postaw propozycjonalnych. W szczególności należałoby prześledzić filozoficzne konsekwencje twierdzenia, że przekonania i pragnienia nie są relacjami między podmiotem a sądami czy treściami. Ponadto broniona tu „ekologiczna" koncepcja psychologii potocznej powinna zostać uzgodniona z wynikami badań empirycznych nad poznaniem społecznym, a w szczególności nad zdolnością czytania umysłów. Postawiłem tu szereg tez na temat natury psychologii potocznej oraz postaw propozycjonalnych. Czy są one prawdopodobne w świetle wiedzy empirycznej na temat praktyki czytania umysłów oraz pojęć postaw propozycjonalnych, jakimi ludzie posługują się w ramach tej praktyki? Ponadto warto zaznaczyć, że także przyjęte tu ujęcie relacji między poziomem osobowym a subosobowym otwiera pole do zadawania nowych, ważkich pytań. Zgodnie z poczynionymi tu ustaleniami przekonania i pragnienia nie wymagają konstytutywnie istnienia jakichś subosobowych odpowiedników. Naturalizacja osobowych stanów intencjonalnych nie wymaga zatem sformułowania koncepcji dotyczącej tego, czym takie subosobowe odpowiedniki miałyby być. Nie znaczy to jednak, że zagadnienie subosobowych podstaw przekonań i pragnień nie jest ważnym problemem *empirycznym* (por. przypis 38, rozdział 5). Nawet jeśli na poziomie subosobowym nie istnieją struktury, które moglibyśmy *zidentyfikować* z przekonaniami i pragnieniami podmiotu, to bez wątpienia coś – jakieś struktury czy procesy – stanowi subosobowy korelat osobowych stanów intencjonalnych. Na przykład dyspozycje konstytuujące posiadanie przekonania, że *p* (należące do dyspozycyjnego stereotypu tego przekonania), mają pewną subosobową bazę kategorialną. Pytanie o naturę tej bazy wydaje się ciekawe, nawet jeśli nie oczekujemy, że istnieje jakiś osobny, pojedynczy (typicznie) stan czy struktura, która byłaby aktywna zawsze i tylko wtedy, gdy osoba działa na podstawie przekonania, że *p*. Być może relacja między poziomami nie polega na *odzwierciedleniu* tego, co

osobowe, w tym, co subosobowe. Niewykluczone jednak, że mamy tu do czynienia z jakąś inną, być może dużo bardziej subtelną i teoretycznie ciekawą relacją.

Mam nadzieję, że powyższe uwagi pokazują wyraźnie, iż zaproponowane w tej książce rozwiązania są płodne, jeśli chodzi o inspirowanie oraz ukierunkowywanie dalszych poszukiwań. Spoglądanie w kierunku roztaczających się perspektyw badawczych nie powinno jednak odwracać naszej uwagi od otrzymanych tu rezultatów. Zaproponowałem koncepcję tego, na czym polega wyjaśnianie reprezentacyjne w kognitywistyce. Zgodnie z nią eksplanacyjnie wartościowe reprezentacje to subosobowe, wewnętrzne modele; reprezentacyjny system poznawczy zaś to system mający u podstaw swojego działania mechanizmy korzystające z takich modeli. Na tej podstawie stwierdziłem, że tezy o śmierci reprezentacjonizmu we współczesnej kognitywistyce są zdecydowanie przedwczesne i nieuzasadnione. Podjąłem się tu wreszcie próby uzgodnienia naukowego oraz manifestującego się obrazu umysłu, pokazując, że błędem jest naturalizowanie osobowych stanów intencjonalnych – ludzkich przekonań, pragnień, intencji i nadziei – przez „szukanie" ich na poziomie subosobowej, mechanistycznej architektury systemu poznawczego.

Spis rysunków i tabel

Rysunki

Rysunek 1. Skonstruowany przez Jamesa Watta regulator kontrolujący działanie silnika parowego 31
Rysunek 2. Hierarchia mechanizmów odpowiadających za zdolność nawigacji przestrzennej u szczurów 75
Rysunek 3. Opisany pierwotnie przez Roberta Cumminsa, w pełni mechaniczny samochód, który przemierza tor, wykorzystując wewnętrzną mapę (S-reprezentację) 208
Rysunek 4. Diagram relacji rodzinnych jako S-reprezentacja 212
Rysunek 5. Schemat przedstawiający wzajemną zależność trzech elementów, które nadają mechanizmowi status MKM 221
Rysunek 6. Podziały w przestrzeni aktywacji drugiej warstwy neuronów w sieci koneksyjnej Cottrella 274
Rysunek 7. Proces emulacji w ujęciu Ricka Grusha 277
Rysunek 8. Samo-modelujący się robot skonstruowany przez Josha Bongarda, Victora Zykova i Hoda Lipsona 286
Rysunek 9. Stephena Sticha i Shauna Nicholsa szkic funkcjonalny procesu podejmowania decyzji 314
Rysunek 10. Poziom osobowy i subosobowy – interpretacja mechanistyczna 341
Rysunek 11. Wirtualny organizm opisany przez Randalla Beera 342
Rysunek 12. Pomiar postaw propozycjonalnych za pomocą skali wyznaczonej przez formy zinterpretowanych wypowiedzi (IUF) w ujęciu Roberta Matthewsa 395

Tabele

Tabela 1. Zestawienie różnic zachodzących między problemem naturalizacji intencjonalności a problemem statusu eksplanacyjnego reprezentacji w kognitywistyce 53

Tabela 2. Williama Ramseya klasyfikacja przednaukowych pojęć reprezentacji 150
Tabela 3. Zestawienie osobowego i subosobowego poziomu wyjaśniania przy mechanistycznej interpretacji dystynkcji osobowe–subosobowe 338

Bibliografia

Adami Christoph (2006), *What do robots dream of?*, „Science", 314, s. 1093–1094.

Anderson John R. (1996), *ACT: A simple theory of complex cognition*, „American Psychologist", 51, s. 355–365.

Anderson Michael, Rosenberg Gregg (2008), *Content and action: The guidance theory of representation*, „Journal of Mind and Behavior", 29, s. 55–86.

Aronson Elliot, Wilson Timothy D., Akert Robin M. (1997), *Psychologia społeczna. Serce i umysł*, przeł. Anna Bezwińska et al., Poznań: Zysk i S-ka.

Atkin Albert (2010), *Peirce's theory of signs*, w: *Stanford Encyclopedia of Philosophy*, ed. Edward Zalta, http://plato.stanford.edu/entries/peirce-semiotics (dostęp: 25.10.2014).

Audi Robert (1994), *Dispositional beliefs and dispositions to believe*, „Noûs", 28, s. 419–434.

Baetu Tudor M. (2012), *Filling in the mechanistic details: Two-variable experiments as tests for constitutive relevance*, „European Journal of Philosophy of Science", 2, s. 337–353.

Baker Lynn R. (1995), *Explaining attitudes. A practical approach to the mind*, Cambridge: Cambridge University Press.

Baker Lynn R. (2001), *Are beliefs brain states?*, w: *Explaining beliefs: Lynne Rudder Baker and her critics*, ed. Anthonie Meijers, Stanford, CA: CSLI Publications, s. 17–38.

Barkow Jerome H., Cosmides Leda, Tooby John, eds. (1992), *The adapted mind: Evolutionary psychology and the generation of culture*, Oxford: Oxford University Press.

Barsalou Lawrence (1999), *Perceptual symbol systems*, „Behavioral and Brain Sciences", 22, s. 577–609.

Barsalou Lawrence (2009), *Simulation, situated conceptualization and prediction*, „Philosophical Transactions of Royal Society B", 364, s. 1281–1289.

Barsalou Lawrence, Solomon Karen O., Wu Ling-Ling (1999), *Perceptual simulation in conceptual tasks*, w: *Cultural, typological, and psycholog-*

ical perspectives in cognitive linguistics: The proceedings of the 4th Conference of the International Cognitive Linguistics Association, Vol. 3, eds. Masako K. Hiraga, Chris Sinha, Sherman Wilcox, Amsterdam: John Benjamins, s. 209–228.

Bartels Andreas (2006), *Defending the structural concept of representation*, „Theoria", 55, s. 7–19.

Bechtel William (1994), *Levels of description and explanation in cognitive science*, „Minds and Machines", 4, s. 1–25.

Bechtel William (1998), *Representations and cognitive explanations: Assesing the dynamicist's challenge in cognitive science*, „Cognitive Science", 22, s. 295–318.

Bechtel William (2008), *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*, London: Routledge.

Bechtel William (2009), *Looking down, around and up: Mechanistic explanation in psychology*, „Philosophical Psychology", 22, s. 543–564.

Bechtel William, Abrahamsen Adele (1993), *Connectionism and the future of folk psychology*, w: *Natural and artificial minds*, ed. Robert G. Burton, Albany, NY: SUNY Press, s. 69–100.

Bechtel William, Abrahamsen Adele (2005), *Explanation: A mechanistic alternative*, „Studies in History and Philosophy of the Biological and Biomedical Sciences", 36, s. 421–441.

Bechtel William, Abrahamsen Adele (2010), *Dynamic mechanistic explanation: Computational modeling of circadian rhythms as an exemplar for cognitive science*, „Studies in History and Philosophy of Science Part A", 41, s. 321–333.

Bechtel William, Mundale Jennifer (1999), *Multiple realizability revisited: Linking cognitive and neural states*, „Philosophy of Science", 66, s. 175––207.

Bechtel William, Richardson Robert C. (1993), *Discovering complexity: Decomposition and localization as strategies in scientific research*, Princeton, NJ: Princeton University Press.

Beer Randall D. (1997), *The dynamics of adaptive behavior: A research program*, „Robotics and Autonomous Systems", 20, s. 257–289.

Beer Randall D. (2003), *The dynamics of active categorical perception in an evolved agent*, „Adaptive Behavior", 11, s. 209–243.

Bennett Max R., Hacker Peter M. S. (2003), *Philosophical foundations of neuroscience*, Oxford: Blackwell.

Bermúdez Jose L. (2000), *Personal and subpersonal: A difference without a distinction*, „Philosophical Explorations", 3, s. 63–82.

Bermúdez Jose L. (2005), *Philosophy of psychology: A contemporary introduction*, London: Routledge.

Bickhard Mark H. (2004a), *The dynamic emergence of representation*, w: *Representation in mind: New approaches to mental representation*, eds. Hugh Clapin, Phillip Staines, Peter Slezak, Oxford: Elsevier Science, s. 71–90.

Bickhard Mark H. (2004b), *Process and emergence: Normative function and representation*, „Axiomathes", 14, s. 135–169.

Bickhard Mark H. (2009), *The interactivist model*, „Synthese", 166, s. 547––591.

Bickle John (2003), *Philosophy and neuroscience: A ruthlessly reductive account*, Dordrecht: Kulwer Academic.

Blachowicz James (1997), *Analog representation beyond mental imagery*, „Journal of Philosophy", 94, s. 55–84.

Block Ned (1986), *Advertisement for a semantics for psychology*, „Midwest Studies in Philosophy", 10, s. 615–678.

Bongard Josh, Zykov Victor, Lipson Hod (2006), *Resilient machines through continuous self-modeling*, „Science", 314, s. 1118–1121.

Bostrom Nick (2003), *Are you living in a computer simulation?*, „Philosophical Quarterly", 53, s. 243–255.

Braddon-Mitchell David, Jackson Frank (2007), *Philosophy of mind and cognition*, Oxford: Blackwell.

Brooks Rodney (1991), *Intelligence without representation*, „Artificial Intelligence", 47, s. 139–159.

Buller David (1993), *Confirmation and the computational paradigm (or: why do you think they call it "artificial" intelligence?)*, „Minds and Machines", 3, s. 155–181.

Busemeyer Jerome R., Townsend James T. (1993), *Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment*, „Psychological Review", 100, s. 432–459.

Calvo Paco (2008), *Towards a general theory of antirepresentationalism*, „British Journal of Philosophy of Science", 59, s. 259–292.

Calvo Paco, García Ángel (2009), *Where is cognitive science heading?*, „Minds and Machines", 19, s. 301–318.

Carroll John W. (2010), *Laws of nature*, w: *Stanford Encyclopedia of Philosophy*, ed. Edward Zalta, http://plato.stanford.edu/entries/laws-of-nature (dostęp: 25.10.2014).

Carruthers Peter (2009), *How we know our own minds: The relationship between mindreading and metacognition*, „Behavioral and Brain Sciences", 32, s. 121–138.

Chalmers David (2010), *Świadomy umysł. W poszukiwaniu teorii fundamentalnej*, przeł. Marcin Miłkowski, Warszawa: PWN.

Chemero Anthony (2009), *Radical embodied cognitive science*, Cambridge, MA: The MIT Press.

Chemero Anthony (2014), *Antyreprezentacjonizm i nastawienie dynamiczne*, przeł. Paweł Gładziejewski, „Przegląd Filozoficzno-Literacki: Kognitywistyka. Reprezentacje", 39, s. 79–107.

Choi Sungho, Fara Michael (2012), *Dispositions*, w: *Stanford Encyclopedia of Philosophy*, ed. Edward Zalta, http://plato.stanford.edu/entries/dispositions (dostęp: 25.10.2014).

Christensen Wayne D., Bickhard Mark H. (2002), *The process dynamics of normative function*, „Monist", 85, s. 3–28.

Churchland Paul M. (1981), *Eliminative materialism and the propositional attitudes*, „Journal of Philosophy", 78, s. 67–90.

Churchland Paul M. (1989), *A neurocomputational perspective: The nature of mind and the structure of science*, Cambridge, MA: The MIT Press.

Churchland Paul M. (2002), *Mechanizm rozumu, siedlisko duszy. Filozoficzna podróż w głąb mózgu*, przeł. Zbigniew Karaś, Warszawa: Aletheia.

Clapin Hugh (2002), *Tacit representation in functional architecture*, w: *Philosophy of mental representation*, ed. Hugh Clapin, Oxford: Oxford University Press, s. 295–311.

Clapin Hugh, Staines Phillip, Slezak Peter, eds. (2004), *Representation in mind: New approaches to mental representation*, Oxford: Elsevier Science.

Clark Andy (1993), *Associative engines*, Cambridge, MA: The MIT Press.

Clark Andy (1997), *Being there: Putting brain, body and world together again*, Cambridge, MA: The MIT Press.

Clark Andy (2013), *Whatever next? Predictive brains, situated agents and the future of cognitive science*, „Behavioral and Brain Sciences", 36, s. 181–204.

Clark Andy, Chalmers David (2008), *Umysł rozszerzony*, przeł. Marcin Miłkowski, w: *Analityczna metafizyka umysłu*, red. Marcin Miłkowski, Robert Poczobut, Warszawa: IFiS PAN, s. 342–357.

Clark Andy, Grush Rick (1999), *Towards a cognitive robotics*, „Adaptive Behavior", 7, s. 5–16.

Clark Andy, Toribio Josefa (1994), *Doing without representing?*, „Synthese", 101, s. 401–431.

Colombo Matteo (2013), *Constitutive relevance and the personal/subpersonal distinction*, „Philosophical Psychology", 26, s. 547–570.

Coltheart Max (2005), *Commentary: conscious experience and delusional belief*, „Philosophy, Psychiatry & Psychology", 12, s. 153–157.
Coltheart Max, Langdon Robyn, McKay Ryan T. (2011), *Delusional belief*, „Annual Review of Psychology", 62, s. 271–298.
Craik Kenneth J. (1943), *The nature of explanation*, Cambridge: Cambridge University Press.
Craver Carl F. (2001), *Role functions, mechanisms, and hierarchy*, „Philosophy of Science", 68, s. 31–55.
Craver Carl F. (2007), *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*, Oxford: Clarendon Press.
Craver Carl F. (2008), *Functions and mechanisms in neuroscience*, w: *Des Neurones à la conscience: Neurophilosophie et philosophie des neurosciences*, éd. Pierre Poirier, Luc Faucher, Eric Racine, Elizabeth Ennan, Bruxelles: De Boeck Université.
Craver Carl F., Bechtel William (2006), *Top-down causation without top-down causes*, „Biology and Philosophy", 22, s. 547–563.
Csibra Gergely (2008), *Goal attribution to inanimate agents by 6.5-month infants*, „Cognition", 107, s. 705–717.
Cummins Robert (1975), *Functional analysis*, „Journal of Philosophy", 72, s. 741–765.
Cummins Robert (1989), *Meaning and mental representation*, Cambridge, MA: The MIT Press.
Cummins Robert (1996), *Representations, targets and attitudes*, Cambridge, MA: The MIT Press.
Cummins Robert (2000), *"How does it work?" versus "What are the laws?": Two conceptions of psychological explanation*, w: *Explanation and cognition*, eds. Frank Keil, Robert A. Wilson, Cambridge, MA: The MIT Press, s. 117–145.
Cummins Robert, Poirier Pierre (2004), *Representation and indication*, w: *Representation in mind: New approaches to mental representation*, eds. Hugh Clapin, Phillip Staines, Peter Slezak, Oxford: Elsevier Science, s. 21–40.
Damasio Antonio (1989), *Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition*, „Cognition", 33, s. 25–62.
Danker Jared F., Anderson John R. (2010), *The ghosts of brain states past: Remembering reactivates the brain regions engaged during encoding*, „Psychological Bulletin", 136, s. 87–102.
Davidson Donald (1963), *Actions, reasons and causes*, „Journal of Philosophy", 60, s. 685–700.

Davidson Donald (1989), *What is present to the mind?*, „Grazer Philosophische Studien", 36, s. 3–18.
Davidson Donald (1992), *Zdarzenia mentalne*, przeł. Tadeusz Baszniak, w: Donald Davidson, *Eseje o prawdzie, języku i umyśle*, red. Barbara Stanosz, Warszawa: PWN, s. 163–193.
Davies Martin (1998), *Language, thought and the language of thought*, w: *Language and thought*, eds. Peter Carruthers, Jill Boucher, Cambridge: Cambridge University Press, s. 226–247.
Davies Martin (2000), *Persons and their underpinnings*, „Philosophical Explorations", 3, s. 43–62.
Dennett Daniel C. (1978), *Brainstorms*, Cambridge, MA: The MIT Press.
Dennett Daniel C. (1995), *Osobowy i subosobowy poziom wyjaśniania: ból*, przeł. Paweł Dziliński, w: *Filozofia umysłu*, red. Bogdan Chwedeńczuk, Warszawa: Aletheia, s. 103–109.
Dennett Daniel C. (1997), *Natura umysłów*, przeł. Witold Turopolski, Warszawa: Wydawnictwo CiS.
Dennett Daniel C. (2003), *Naprawdę przekonani: strategia intencjonalna i dlaczego ona działa*, przeł. Marcin Miłkowski, „Przegląd Filozoficzno-Literacki", 6, s. 87–109.
Dennett Daniel C. (2008), *Rzeczywiste wzorce*, przeł. Marcin Miłkowski, w: *Analityczna metafizyka umysłu. Najnowsze kontrowersje*, red. Marcin Miłkowski, Robert Poczobut, IFiS PAN: Warszawa, s. 299–325.
Drayson Zoe (2012), *The uses and abuses of personal/subpersonal distinction*, „Philosophical Perspectives", 26, s. 1–18.
Dretske Fred (1981), *Knowledge and the flow of information*, Cambridge, MA: The MIT Press.
Dretske Fred (1986), *Misrepresentation*, w: *Belief*, ed. Radu Bogdan, Oxford: Oxford University Press, s. 17–36.
Dretske Fred (1988), *Explaining behavior: Reasons in the world of causes*, Cambridge, MA: The MIT Press.
Dreyfus Hubert (1972), *What computers can't do*, New York, NY: Harper and Row.
Eckhardt Barbara von (1993), *What is cognitive science?*, Cambridge, MA: The MIT Press.
Egan Frances (1995), *Folk psychology and cognitive architecture*, „Philosophy of Science", 62, s. 179–196.
Fodor Jerry (1975), *The Language of thought*, Cambridge, MA: Harvard University Press.
Fodor Jerry (1981), *RePresentations: Philosophical essays on the foundations of cognitive science*, Cambridge, MA: The MIT Press.

Fodor Jerry (1987), *Psychosemantics: The problem of meaning in the philosophy of mind*, Cambridge, MA: The MIT Press.
Fodor Jerry (1990), *A theory of content and other essays*, Cambridge, MA: The MIT Press.
Fodor Jerry (2001), *Eksperci od wiązów. Język myśleński i jego semantyka*, przeł. Marcin Gokieli, Warszawa: Aletheia.
Fodor Jerry (2008), *Nauki szczegółowe (albo: niejednorodność nauki jako hipoteza robocza)*, przeł. Marcin Gokieli, w: *Analityczna metafizyka umysłu. Najnowsze kontrowersje*, red. Marcin Miłkowski, Robert Poczobut, Warszawa: IFiS PAN, s. 56–75.
Freeman Walter J., Skarda Christine A. (1990), *Representations: who needs them?*, w: *Brain organization and memory: Cells, systems and circuits*, eds. James L. McGaugh, Norman M. Weinberger, Gary Lynch, New York, NY: Oxford University Press, s. 375–380.
Friston Karl J., Stephan Klass E. (2007), *Free-energy and the brain*, „Synthese", 159, s. 417–458.
Gallagher Shaun (2005), *How the body shapes the mind*, Oxford: Oxford University Press.
García Ángel, Calvo Paco (2010), *Is cognition a matter of representations? Emulation, teleology, and time-keeping in biological systems*, „Adaptive Behavior", 18, s. 400–415.
Gardenförs Peter (1996), *Cued and detached representations in animal cognition*, „Behavioural Processes", 35, s. 263–273.
Gardenförs Peter (2000), *Conceptual spaces: The geometry of thought*, Cambridge, MA: The MIT Press.
Gardenförs Peter (2007), *Mind-reading as control theory*, „European Review", 15, s. 223–240.
Gelder Tim van (1995), *What might cognition be if not computation?*, „Journal of Philosophy", 91, s. 345–381.
Gergely György, Csibra Gergely (2003), *Teleological reasoning in infancy: naive theory of rational action*, „Trends in Cognitive Sciences", 7, s. 287–292.
Gerrans Philip (2002), *The theory of mind module in evolutionary psychology*, „Biology and Philosophy", 17, s. 305–321.
Gibson James J. (1979), *The ecological approach to visual perception*, Boston, MA: Houghton Mifflin.
Giere Ronald N. (2004), *How models are used to represent reality*, „Philosophy of Science", 71, s. 742–752.
Giere Ronald N. (2010), *An agent-based conception of models and scientific representation*, „Synthese", 172, s. 269–281.

Gładziejewski Paweł (2012), *William Ramsey o psychologii potocznej, racjonalności i pojęciu reprezentacji w naukach kognitywnych*, „Diametros", 31, s. 33–55.

Gładziejewski Paweł (2013a), *Reprezentacjonizm a wyjaśnianie mechanistyczne w kognitywistyce*, „Filozofia Nauki", 4, s. 51–77.

Gładziejewski Paweł (2013b), *Shared representations, perceptual symbols and the vehicles of mental concepts*, „Journal of Consciousness Studies", 20, s. 102–123.

Gładziejewski Paweł (2015), *Action guidance is not enough, representations need correspondence too: A plea for a two-factor theory of representation*, „New Ideas in Psychology", doi:10.1016/j.newideapsych.2015.01.005.

Gładziejewski Paweł (w druku), *Explaining mental phenomena with internal representations. A mechanistic perspective*, „Studies in Logic, Grammar and Rhetoric".

Glennan Stuart (2002), *Rethinking mechanistic explanation*, „Philosophy of Science", 69, s. S342–S353.

Godfrey-Smith Peter (2002), *Environmental complexity and the evolution of cognition*, w: *The evolution of intelligence*, eds. Robert Sternberg, James Kaufman, Mahwah, NJ: Lawrence Erlbaum Press, s. 233–249.

Godfrey-Smith Peter (2004), *On folk psychology and mental representation*, w: *Representation in mind: New approaches to mental representation*, eds. Hugh Clapin, Phillip Staines, Peter Slezak, Oxford: Elsevier Science, s. 147–162.

Godfrey-Smith Peter (2005), *Reduction in real life*, w: *Being reduced: New essays on reduction, explanation, and causation*, eds. Jakob Hohwy, Jesper Kallestrup, Oxford: Oxford University Press, s. 52–74.

Godfrey-Smith Peter (2006), *The strategy of model-based science*, „Biology and Philosophy", 21, s. 725–740.

Gold Ian, Stoljar Daniel (1999), *A neuron doctrine in the philosophy of neuroscience*, „Behavioral and Brain Sciences", 22, s. 809–830.

Goldman Alvin (2006), *Simulating minds: The philosophy, psychology and neuroscience of mindreading*, Oxford: Oxford University Press.

Goldman Alvin (2007), *Philosophical intuitions: their target, their source, and their epistemic status*, „Grazer Philosophische Studien", 74, s. 1–26.

Goldman Alvin (2012), *A moderate approach to embodied cognitive science*, „Review of Philosophy and Psychology", 3, s. 71–88.

Golomb Beatrice, Sejnowski Terrence J. (1995), *Sex recognition from faces using neural networks*, w: *Applications of neural networks*, ed. Alan Murray, Kulwer Academic Publishers, s. 71–92.

Gorman Paul R., Sejnowski Terrence J. (1988), *Analysis of the hidden units in a layered network trained to classify sonar targets*, „Neural Networks", 1, s. 75–89.

Grobler Adam (2006), *Metodologia nauk*, Kraków: Znak.

Grush Rick (1997), *The architecture of representation*, „Philosophical Psychology", 10, s. 5–23.

Grush Rick (2004), *The emulation theory of representation: motor control, imagery and perception*, „Behavioral and Brain Sciences", 27, s. 377––442.

Grush Rick (2008), *Review of "Representation Reconsidered" by W. Ramsey*, „Notre Dame Philosophical Reviews", http://ndpr.nd.edu/news/23327/?id=12243 (dostęp: 25.10.2014).

Grush Rick (2010), *Emulujący wywiad z... Rickiem Grushem*, przeł. Jakub Matyja, Piotr Momot, „Avant", 1, s. 199–211.

Gulick Robert van (2008), *Redukcja, emergencja i inne nowsze stanowiska na temat problemu umysł–ciało. Przegląd filozoficzny*, przeł. Robert Poczobut, w: *Analityczna metafizyka umysłu. Najnowsze kontrowersje*, red. Marcin Miłkowski, Robert Poczobut, Warszawa: IFiS PAN, s. 144––190.

Hacking Ian (1983), *Representing and intervening: Introductory topics in the philosophy of natural science*, Cambridge: Cambridge University Press.

Harman Gilbert (1987), *(Non-solipsistic) conceptual role semantics*, w: *New directions in semantics*, ed. Ernest LePore, London: Academic Press, s. 55–81.

Haselager Pim, Groot Andre de, Rappard Hans van (2003), *Representationalism vs. anti-representationalism: A debate for the sake of appearance*, „Philosophical Psychology", 16, s. 5–23.

Haugeland John (1998), *Representational genera*, w: John Haugeland, *Having thought. Essays in the metaphysics of mind*, Cambridge, MA: Harvard University Press, s. 171–206.

He Zijing, Bolz Matthias, Baillargeon Renée (2011), *False-belief understanding in 2.5-years olds: evidence from violation-of-expectation change-of-location and unexpected-content tasks*, „Developmental Science", 14, s. 292–305.

Hempel Carl, Oppenheim Paul (1948), *Studies in the logic of explanation*, „Philosophy of Science", 15, s. 135–175.

Henderson David, Horgan Terrence (2004), *What does it take to be a true believer? Against the opulent ideology of eliminative materialism*, w: *Mind as a scientific object: Between brain and culture*, eds. Christi-

na E. Erneling, David M. Johnson, Oxford: Oxford University Press, s. 211–224.
Herschbach Mitchell (2008), *Folk psychological and phenomenological accounts of social perception*, „Philosophical Explorations", 11, s. 223––235.
Herschbach Mitchell (2012), *Mirroring versus simulation: on the representational function of simulation*, „Synthese", 189, s. 483–513.
Hippel William von, Trivers Robert (2011), *The evolution and psychology of self-deception*, „Behavioral and Brain Sciences", 34, s. 1–16.
Hohwy Jakob (2013), *The predictive mind*, Oxford: Oxford University Press.
Horgan Terrence (1993), *The austere ideology of folk psychology*, „Mind & Language", 8, s. 282–297.
Horgan Terrence, Graham George (1991), *In defense of southern fundamentalism*, „Philosophical Studies", 62, s. 107–134.
Hornsby Jennifer (2000), *Personal and sub-personal: A defense of Dennett's early distinction*, „Philosophical Explorations", 3, s. 6–24.
Hurley Susan (2008), *The shared circuits model. How control, mirroring, and simulation can enable deliberation, imitation and mindreading*, „Behavioral and Brain Sciences", 31, s. 1–22.
Hutchins Edwin (1995), *Cognition in the wild*, Cambridge, MA: The MIT Press.
Illari Phyllis (2013), *Mechanistic explanation: integrating the ontic and epistemic*, „Erkenntnis", 78, s. 237–255.
Johnson-Laird Philip N. (1983), *Mental models: Towards a cognitive science of language, inference and consciousness*, Cambridge, MA: Harvard University Press.
Johnson-Laird Philip N. (2005), *The history of mental models*, w: *Psychology of reasoning: Theoretical and historical perspectives*, eds. Ken Manktelow, Man C. Chung, London: Psychology Press, s. 179–212.
Kaplan David M. (2011), *Explanation and description in computational neuroscience*, „Synthese", 183, s. 339–373.
Kaplan David M., Bechtel William (2009), *Dynamical models: an alternative or complement to mechanistic explanations?*, „Topics in Cognitive Science", 3, s. 438–444.
Kent Christopher, Lamberts Koen (2008), *The encoding-retrieval relationship: Retrieval as mental simulation*, „Trends in Cognitive Sciences", 12, s. 92–98.
Kim Jaegwon (2002), *Umysł w świecie fizycznym*, przeł. Robert Poczobut, Warszawa: IFiS PAN.

Kirsh David, Maglio Paul (1994), *On distinguishing epistemic from pragmatic action*, „Cognitive Science", 18, s. 513–549.

Kitcher Philip (1989), *Explanatory unification and the causal structure of the world*, w: *Scientific explanation*, eds. Philip Kitcher, Wesley Salmon, Minneapolis, MN: University of Minnesota Press, s. 410–505.

Laird John E., Newell Allen, Rosenbloom Paul S. (1987), *SOAR: An architecture for general intelligence*, „Artificial Intelligence", 33, s. 1–64.

Lettvin Jerome, Maturana Humberto, McCulloch Warren, Pitts Walter (1959), *What the frog's eye tells the frog's brain*, „Proceedings of the Institute of Radio Engineers", 47, s. 1940–1951.

Leuridan Bert (2012), *Three problems for the mutual manipulability account of constitutive relevance in mechanisms*, „British Journal for the Philosophy of Science", 63, s. 399–427.

Logothetis Nikos K., Sheinberg David L. (1996), *Visual object recognition*, „Annual Review of Neuroscience", 19, s. 577–621.

Looren de Jong Huib, Schouten Maurice K. D. (2005), *Ruthless reductionism: A review essay of John Bickle's "Philosophy and neuroscience: A ruthlessly reductive account"*, „Philosophical Psychology", 18, s. 473–486.

Lycan William (2008), *Phenomenal intentionalities*, „American Philosophical Quarterly", 45, s. 233–252.

Machamer Peter, Darden Lindsey, Craver Carl F. (2011), *Myślenie w kategoriach mechanizmów*, przeł. Witold Hensel, „Przegląd Filozoficzno-Literacki: Filozofia Biologii", 31, s. 145–176.

Machery Edouard (2007), *Concept empiricism: a methodological critique*, „Cognition", 104, s. 19–46.

Mameli Matteo (2001), *Mindreading, mindshaping, and evolution*, „Biology and Philosophy", 16, s. 597–628.

Marr David (2010), *Vision: A computational investigation into the human representation and processing of visual information*, Cambridge, MA: The MIT Press.

Martin Alex (2007), *The representation of object concepts in the brain*, „Annual Review of Psychology", 58, s. 25–45.

Matthews Robert J. (2007), *The measure of mind: Propositional attitudes and their attribution*, Oxford: Oxford University Press.

Matthews Robert J. (2011), *Measurement-theoretic accounts of propositional attitudes*, „Philosophy Compass", 6, s. 828–841.

McCauley Robert N. (1986), *Intertheoretic relations and the future of psychology*, „Philosophy of Science", 53, s. 179–199.

McCauley Robert N., Bechtel William (2001), *Explanatory pluralism and the heuristic identity theory*, „Theory and Psychology", 11, s. 736–760.

McDowell John (1994), *The content of perceptual experience*, „Philosophical Quarterly", 44, s. 190–205.

McKitrick Jennifer (2005), *Are dispositions causally relevant?*, „Synthese", 144, s. 357–371.

Metzinger Thomas (2003), *Being no one. The self-model theory of subjectivity*, Cambridge, MA: The MIT Press.

Meyer Kaspar, Damasio Antonio (2009), *Convergence and divergence in a neural architecture for recognition and memory*, „Trends in Neurosciences", 32, s. 376–382.

Miłkowski Marcin (2013), *Explaining the computational mind*, Cambridge, MA: The MIT Press.

Miłkowski Marcin, Poczobut Robert (2005), *Czym jest i jak istnieje umysł?*, „Diametros", 3, s. 27–55.

Millikan Ruth G. (1984), *Language, thought and other biological categories*, Cambridge: Cambridge University Press.

Millikan Ruth G. (2002), *Varieties of meaning*, Cambridge, MA: The MIT Press.

Morgan Alex (2014), *Representations gone mental*, „Synthese", 191, s. 213––244.

Nagel Ernst (1970), *Struktura nauki. Zagadnienia logiki wyjaśnień naukowych*, przeł. Jerzy Giedymin, Bożydar Rassalski, Helena Eilstein, Warszawa: PWN.

Newell Allen (1980), *Physical symbol systems*, „Cognitive Science", 4, s. 135––183.

Niedenthal Paula M., Winkielman Piotr, Mondillon Laurlie, Vermeulen Nicolas (2009), *Embodiment of emotional concepts: Evidence from EMG measures*, „Journal of Personality and Social Psychology", 96, s. 1120––1136.

Noë Alva (2002), *Is the visual world a grand illusion?*, „Journal of Consciousness Studies", 9, s. 1–12.

Nowakowski Przemysław (2010), *Fantom ciała jako cielesna samoświadomość*, „Avant", 1, s. 225–246.

O'Brien Gerard, Opie Jon (2004), *Notes toward a structuralist theory of mental representation*, w: *Representation in mind: New approaches to mental representation*, eds. Hugh Clapin, Phillip Staines, Peter Slezak, Oxford: Elsevier Science, s. 1–20.

O'Regan Kevin J., Noë Alva (2008), *Sensomotoryczne ujęcie widzenia i świadomości wzrokowej*, w: *Formy aktywności umysłu: ujęcia kognitywistyczne*, red. Andrzej Klawiter, t. 1: *Emocje, percepcja, świadomość*, Warszawa: PWN, s. 138–236.

Onishi Kristine H., Baillargeon Renée (2005), *Do 15-month-old infants understand false beliefs?*, „Science", 308, s. 255–258.

Oppenheim Paul, Putnam Hilary (1958), *Unity of science as a working hypothesis*, w: *Minnesota studies in the philosophy of sicence*, Vol. 2: *Concepts, theories and the mind-body problem*, eds. Herbert Feigl, Michael Scriven, Grover Maxwell, Minneapolis, MN: University of Minnesota Press, s. 3–36.

Palmer Stephen (1979), *Fundamental aspects of cognitive representation*, w: *Cognition and categorization*, eds. Eleanor Rosch, Barbara B. Lloyd, Hillsdale, MI: Lawrence Erblaum, s. 259–303.

Papineau David (1987), *Reality and representation*, Oxford: Blackwell.

Peacocke Christopher (1983), *Sense and content*, Oxford: Oxford University Press.

Pecher Diane, Zeelenberg Rene, Barsalou Lawrence (2003), *Verifying properties from different modalities for concepts produces switching costs*, „Psychological Science", 14, s. 119–124.

Peirce Charles S. (1997), *Wybór pism semiotycznych*, oprac. Hanna Buczyńska-Garewicz, przeł. Ryszard Mirek, Andrzej J. Nowak, Warszawa: Polskie Towarzystwo Semiotyczne.

Pezzulo Giovanni (2011), G*rounding procedural and declarative knowledge in sensorimotor anticipation*, „Mind & Language", 26, s. 78–114.

Piccinini Gualtiero (2007a), *Computational explanation and mechanistic explanation of mind*, w: *Cartographies of the mind: The interface between philosophy and cognitive science*, eds. Mario de Caro, Francesco Ferretti, Massimo Marraffa, Dordrecht: Springer, s. 23–36.

Piccinini Gualtiero (2007b), *Computing mechanisms*, „Philosophy of Science", 74, s. 501–526.

Piccinini Gualtiero (2008), *Computation without representation*, „Philoso-phical Studies", 137, s. 205–241.

Piccinini Gualtiero, Craver Carl F. (2011), *Integrating psychology and neuroscience: functional analyses as mechanism sketches*, „Synthese", 183, s. 283–311.

Piłat Robert (1999), *Umysł jako model świata*, Warszawa: IFiS PAN.

Piłat Robert (2006), *Symulacja jako mechanizm percepcji*, w: Robert Piłat, *Doświadczenie i pojęcie. Studia z fenomenologii i filozofii umysłu*, Warszawa: IFiS PAN, s. 61–78.

Pinedo Manuel de, Noble Jason (2008), *Beyond persons: extending the personal/subpersonal distinction to non-rational and artificial agents*, „Biology and Philosophy", 23, s. 87–100.

Poczobut Robert (2009), *Między redukcją a emergencją. Spór o miejsce umysłu w świecie fizycznym*, Wrocław: UWr.

Prinz Jesse (2002), *Furnishing the mind: Concepts and their perceptual basis*, Cambridge, MA: The MIT Press.

Prinz Jesse (2010), *Can concept empiricism forestall concept eliminativism?*, „Mind & Language", 25, s. 612–621.

Putnam Hilary (1963), *Brains and behavior*, w: *Analytical philosophy*, ed. Ronald Butler, Oxford: Basil Blackwell, s. 211–235.

Quiroga Quian, Reddy Leila, Kreiman Gabriel, Koch Christof, Fried Itzhak (2005), *Invariant visual representation by single neurons in the human brain*, „Nature", 435, s. 1102–1107.

Raatikainen Panu (2010), *Causation, exclusion, and the special sciences*, „Erkenntnis", 73, s. 349–363.

Raatikainen Panu (2013), *Can the mental be causally efficacious?*, w: *Regarding the mind, naturally: Naturalist approaches to the sciences of the mental*, eds. Konrad Talmont-Kamiński, Marcin Miłkowski, Newcastle upon Tyne: Cambridge Scholars Publishing, s. 138–166.

Ramsey William (2007), *Representation Reconsidered*, Cambridge: Cambridge University Press.

Ramsey William, Stich Stephen, Garon Joseph (1990), *Connectionism, eliminativism and the future of folk psychology*, „Philosophical Perspectives", 4, s. 499–533.

Ratcliffe Matthew (2007), *Rethinking commonsense psychology. A crtitique of folk psychology, theory of mind and simulation*, Basingstoke: Pellgrave Macmillan.

Robbins Philip, Aydede Murat (2008), *The cambridge handbook of situated cognition*, Cambridge: Cambridge University Press.

Rogers Timothy T., McClelland John (2004), *Semantic cognition: A parallel distributed processing approach*, Cambridge, MA: The MIT Press.

Rooij Iris van, Bongers Raoul M., Haselager Pim (2002), *A non-representational approach to imagined action*, „Cognitive Science", 26, s. 345–375.

Rupert Robert (2009), *Cognitive systems and the extended mind*, New York, NY: Oxford University Press.

Ryder Dan (2004), *SINBaD neurosemantics: a theory of mental representation*, „Mind & Language", 19, s. 211–240.

Ryle Gilbert (1970), *Czym jest umysł?*, przeł. Witold Marciszewski, Warszawa: PWN.

Salmon Wesley (1971), *Statistical explanation*, w: *Statistical explanation and statistical relevance*, ed. Wesley Salmon, Pittsburgh, PA: University of Pittsburgh Press, s. 29–87.

Salmon Wesley (1984), *Scientific explanation and the causal structure of the world*, Princeton, NJ: Princeton University Press.

Schacter Daniel L., Addis Donna R., Hassabis Demis, Martin Victoria C., Spreng R. Nathan, Szpunar Karl K. (2012), *The future of memory: Remembering, imagining, and the brain*, „Neuron", 76, s. 677–694.

Schindler Samuel (2013), *Mechanistic explanation: asymmetry lost*, w: *Recent progress in philosophy of science: Perspectives and foundational problems. The third european philosophy of science association proceedings*, eds. Vassilios Karakostas, Dennis Dieks, Dordrecht: Springer, s. 81–91.

Schwitzgebel Eric (2001), *In-between believing*, „Philosophical Quarterly", 51, s. 76–82.

Schwitzgebel Eric (2002), *A phenomenal, dispositional account of belief*, „Noûs", 36, s. 249–275.

Schwitzgebel Eric (2007), D*o you have constant tactile experience of your feet in your shoes? Or is experience limited to what's in attention?*, „Journal of Consciousness Studies", 14, s. 5–35.

Schwitzgebel Eric (2010a), *Acting contrary to our professed beliefs, or: the gulf between occurrent judgment and dispositional belief*, „Pacific Philosophical Quartlerly", 91, s. 531–551.

Schwitzgebel Eric (2010b), *Belief*, w: *Stanford Encyclopedia of Philosophy*, ed. Edward Zalta, http://plato.stanford.edu/entries/belief (dostęp: 25.10.2014).

Schwitzgebel Eric (2013), *A dispositional approach to the attitudes: thinking outside of the belief box*, w: *New essays on belief: Structure, constitution and content*, ed. Nikolaj Nottelmann, London: Palgrave Macmillan, s. 75–100.

Searle John (1980), *Minds, brains and programs*, „Behavioral and Brain Sciences", 3, s. 417–457.

Sejnowski Terrence J., Rosenberg Charles R. (1987), *Parallel networks that learn to pronounce English text*, „Complex Systems", 1, s. 145–168.

Sellars Wilfried (1963), *Philosophy and the scientific image of man*, w: Wilfried Sellars, *Science, perception and reality*, London: Routledge & Kegan Paul, s. 1–40.

Shagrir Oron (2010), *Brains as analog-model computers*, „Studies in History and Philosophy of Science", 41, s. 271–279.

Shagrir Oron (2012), *Structural representations in the brain*, „British Journal for the Philosophy of Science", 63, s. 519–545.

Shapiro Lawrence (2000), *Multiple realizations*, „Journal of Philosophy", 97, s. 635–654.

Shapiro Lawrence (2010), *Embodied cognition*, New York, NY: Routledge Press.

Shapiro Lawrence (2011), *Mental manipulations and the problem of causal exclusion*, „Australasian Journal of Philosophy", 90, s. 507–524.

Shoemaker Sydney (1980), *Causality and properties*, w: *Time and cause*, ed. Peter van Inwagen, Dordrecht: Reidel, s. 109–135.

Short Thomas L. (2007), *Peirce's theory of signs*, Cambridge: Cambridge University Press.

Simmons Kyle W., Ramjee Vimal, Beauchamp Michael S., McRae Ken, Martin Alex, Barsalou Lawrence (2007), *A common neural substrate for perceiving and knowing about color*, „Neuropsychologia", 45, s. 2802–2810.

Skidelsky Liza (2006), *Personal-subpersonal: the problem of interlevel relations*, „Protosociology", 22, s. 120–139.

Sloman Aaron (2011), *Comments on "The Emulating Interview... with Rick Grush"*, „Avant", 2, s. 35–44.

Solomon Karen O., Barsalou Lawrence (2004), *Perceptual simulation in property verification*, „Memory & Cognition", 32, s. 119–140.

Spaulding Shannon (2010), *Embodied cognition and mindreading*, „Mind & Language", 25, s. 579–599.

Sprevak Mark (2011), *Review of William M. Ramsey, "Representation Reconsidered"*, „British Journal for the Philosophy of Science", 62, s. 669–675.

Sterelny Kim (2003), *Thought in a hostile world: The evolution of human cognition*, New York, NY: Wiley-Blackwell.

Stewart John, Gapenne Olivier, Di Paolo Ezequiel A., eds. (2010), *Enaction: Toward a new paradigm for cognitive science*, Cambridge, MA: Bradford Books.

Stich Stephen (1983), *From folk psychology to cognitive science*, Cambridge, MA: The MIT Press.

Stich Stephen (1992), *What is a theory of mental representation?*, „Mind", 101, s. 243–261.

Stich Stephen, Nichols Shaun (1992), *Folk psychology: simulation or tacit theory?*, „Mind & Language", 7, s. 35–71.

Strawson Galen (1994), *Mental reality*, Cambridge, MA: MIT Press.

Suárez Mauricio (2003), *Scientific representation: against similarity and isomorphism*, „International Studies in the Philosophy of Science", 17, s. 225–244.

Swoyer Christopher (1991), *Structural representation and surrogative reasoning*, „Synthese", 87, s. 449–508.

Świątczak Bartłomiej (2003), *Przekonania jako przyczyny zachowań. Dyskusja z koncepcją Freda Dretskego*, „Przegląd Filozoficzny – Nowa Seria", 48, s. 61–77.

Tabery James (2004), *Synthesizing activities and interactions in the concept of a mechanism*, „Philosophy of Science", 71, s. 1–15.

Thagard Paul (2010), *Cognitive science*, w: *Stanford Encyclopedia of Philosophy*, ed. Edward Zalta, http://plato.stanford.edu/entries/cognitive-science (dostęp: 25.10.2014).

Thelen Esther, Smith Linda B. (1994), *A dynamic systems theory approach to the development of cognition and action*, Cambridge, MA: The MIT Press.

Thompson Evan (2007), *Mind in life: Biology, phenomenology, and the sciences of mind*, Cambridge, MA: The Belknap Press.

Thompson Evan, Rosch Eleanor (1992), *The embodied mind: Cognitive science and human experience*, Cambridge, MA: The MIT Press.

Tinbergen Nikolas (1963), *On aims and methods of ethology*, „Zeitschrift für Tierpsychologie", 20, s. 410–433.

Toribio Josefa (1998), *Meaning and other non-biological categories*, „Philosophical Papers", 27, s. 129–150.

Weiskopf Daniel (2007), *Concept empiricism and the vehicles of thought*, "Journal of Consciousness Studies", 14, s. 156–183.

Whiten Andrew (1997), *The Machiavellian mindreader*, w: *Machiavellian intelligence II: Extensions and evaluations*, eds. Andrew Whiten, Richard Byrne, Cambridge: Cambridge University Press, s. 144–173.

Wilson Robert A., Foglia Lucia (2011), *Embodied cognition*, w: *Stanford Encyclopedia of Philosophy*, ed. Edward Zalta, http://plato.stanford.edu/entries/embodied-cognition (dostęp: 25.10.2014).

Wimsatt William (1976), *Reductionism, levels of organization, and the mind-body problem*, w: *Consciousness and the brain*, eds. Gordon Globus, Irwin Savodnik, Grover Maxwell, New York, NY: Plenum Press, s. 199––267.

Wimsatt William (2006a), *Aggregate, composed, and evolved systems: Reductionistic heuristics as means to more holistic theories*, „Biology and Philosophy", 21, s. 667–702.

Wimsatt William (2006b), *Reductionism and its heuristics: Making methodological reductionism honest*, „Synthese", 151, s. 445–475.

Wittgenstein Ludwig (2000), *Dociekania filozoficzne*, przeł. Bogusław Wolniewicz, Warszawa: PWN.

Woodward James (2003), *Making things happen: A theory of causal explanation*, Oxford: Oxford University Press.

Woodward James (2008a), *Causation and manipulability*, w: *Stanford Encyclopedia of Philosophy*, ed. Edward Zalta, http://plato.stanford.edu/entries/causation-mani (dostęp: 25.10.2014).

Woodward James (2008b), *Mental causation and neural mechanisms*, w: *Being reduced: New essays on reduction, explanation, and causation*, eds. Jakob Hohwy, Jesper Kallestrup, Oxford: Oxford University Press, s. 218–262.

Woodward James (2010), *Causation in biology: stability, specificity, and the choice of levels of explanation*, „Biology and Philosophy", 25, s. 287–318.

Wright Cory D. (2012), *Mechanistic explanation without the ontic conception*, „European Journal of Philosophy of Science", 2, s. 375–394.

Wu Ling-Ling, Barsalou Lawrence (2009), *Perceptual simulation in conceptual combination: Evidence from property generation*, „Acta Psychologica", 132, s. 173–189.

Yablo Stephen (1992), *Mental causation*, „Philosophical Review", 101, s. 245–280.

Zahavi Dan (2006), *Subjectivity and selfhood: Investigating the first-person perspective*, Cambridge, MA: The MIT Press.

Zawidzki Tadeusz (2013), *Mindshaping. A new framework for understanding human social cognition*, Cambridge, MA: The MIT Press.

Zednik Carlos (2011), *The nature of dynamical explanation*, „Philosophy of Science", 78, s. 238–263.

Żegleń Urszula (2003), *Filozofia umysłu. Dyskusja z naturalistycznymi koncepcjami umysłu*, Toruń: Adam Marszałek.

Summary
Explaining with mental representations. A mechanistic perspective

It has been traditionally assumed that mental representations are a crucial part of the explanatory repertoire of cognitive science. The mind, according to this traditional view, works by internally representing the world; thus when trying to explain its workings one should invoke representations. However, recently the notion of internal representation has become highly controversial in at least two senses. Firstly, whether internal representations are indeed useful in the project of providing scientific explanations of cognitive phenomena is hotly debated, and it seems like the tide has been steadily turning in favor of the antirepresentationalist position. Secondly, there is another problem – one which, unfortunately, is not recognized in the literature as often as it should – in that it seems like philosophers and cognitive scientists lack a well-developed, universally agreed upon concept of what representational explanations are in the first place. That is, there is no principled answer to the question of what criteria should a given cognitive-scientific explanation meet in order to count as (truly, genuinely, nontrivially, etc.) representational.

It is the latter problem – the problem of the nature of representational explanation in cognitive science – that constitutes the main subject matter of the present book. My aim is to develop and defend the view of what cognitive-scientific representational explanation amounts to, and then to draw some consequences from such a view. Throughout the book, I assume that the basic form of explanation in cognitive science is mechanistic explanation, i.e. that cognitive scientists explain phenomena by showing how they are enabled by their mechanisms, construed as organized, functionally characterized component parts of cognitive systems.

The book consists of five chapters. The first two chapters are largely introductory and serve to lay conceptual and theoretical foundations for subsequent ones. In the first chapter, I describe recent debates over the explanatory status of mental representations in cognitive science. It is there that I also introduce the distinction between the first- and second-order problem of representational explanation, where the former pertains to whether representations are explanatorily useful for cognitive science, and the latter pertains to the question of what criteria should an explanation meet to count as genuinely representational at all. I also pose a problem of how two different philosophical projects are related, namely, on the one hand, the project of showing how representational explanation in cognitive science should be construed, and, on the other hand, the project of naturalizing personal-level intentional states like beliefs, desires, intentions and other propositional attitudes.

In the second chapter, I present the mechanistic view of explanation, in particular with extension to explanation in cognitive science. I show why and in what sense (a large portion of) cognitive-scientific explanatory practice should be construed as consisting in discovering and describing mechanisms that underlie cognition.

In the third chapter, I attempt to apply the mechanistic model of explanation to what I have earlier dubbed the "second-order" problem of representational explanation. I turn the general question of what representational explanations are into a more precise and tractable question of what mechanistic representational explanations are. I define the latter as explanations that posit a specific kind of mechanism, namely a mechanism that makes use of component(s) whose operation (function) consists in representing something. I argue that a satisfying theory of representational mechanisms should meet what William Ramsey (2007) calls the "job description" challenge. This means that for a mechanism component to count as a representation – and thus for a larger mechanism that contains this component to count as representational – it needs to be possible to show how or in what sense exactly this component performs a function that is recognizably and nontrivially representational.

In the fourth chapter, I lay out the main thesis of the book by proposing a solution to the second-order problem of representational explanation in cognitive science. I argue that a cognitive-scientific explanation is representational if it consists in explaining a given phenomenon by a mechanism equipped with a "consumable" (i.e. potentially exploitable) structural model of a given domain. I present a functional sketch of a mechanism of this kind. On this view, a representational mechanism has a component part whose proper functioning nonacidentally depends (i.e. depends so due to the functional organization of the mechanism) on whether a relation of structural similarity holds between this component itself and some other (represented) domain. Thus, according to the proposal defended in the book, representational mechanisms work by employing internal models-surrogates of (fragments of) reality. I also defend my proposal against a number of possible objections. Among other things, I attempt to show that my view is immune to arguments traditionally raised against similarity-based accounts of representation, as well as defend the legitimacy of the theoretical distinction between detector (i.e. covariance-based) and structural representations (i.e. models).

In the same chapter, I try to draw from the abovementioned view some consequences for the first-order problem of representational explanation, namely the problem of whether representations are in fact an useful explanatory tool in the sciences of the mind. In particular, I concentrate on the role played by representations in what may be called the post-classical cognitive science, i.e. in theoretical approaches that rose to prominence only after the classical approach – which viewed cognition as explicit-rule-based symbolic computation – had lost its paradigmatic status. Basing on my theory of representational mechanisms, I make an assumption that representations play an explanatory role in post-classical cognitive science to the extent that the latter postulates, as explanantia, mechanisms that make use of internal, exploitable models. Using some examples from actual cognitive science, I show that the mechanisms which meet my proposed criteria for being representational play a major explanatory role in a number

of post-classical approaches, including connectionism, control theory, cognitive robotics, as well as cognitive and computational neuroscience. I conclude that the case for representationalism in (modern) cognitive science is as strong as it has ever been.

The fifth and final chapter is devoted to critically examining the relation between the question of representational explanation in cognitive science and the project of naturalizing intentionality, that is, the project of, broadly speaking, finding a place in the natural order for beliefs, desires and other propositional attitudes. Once again, I employ the mechanistic model of explanation as a theoretical guide. I argue that there are no strong philosophical grounds for thinking, as many contemporary authors do, that folk-psychological explanations which invoke propositional attitudes are (convoluted) mechanistic explanations or that they have specific commitments with regards to the internal, mechanistic architecture of cognition. I claim that in this particular sense, the practice of explaining actions by attributing beliefs, desires and other personal-level states does not require to be legitimized by the (subpersonal) facts regarding the mechanistic organization of cognitive systems. I argue that properly understood folk psychological explanations concentrate on the highest, "ecological" or systemic level of mechanistic organization and work by connecting the cognitive system as a whole to its social and natural environment. This enables me to show how we can claim that although folk-psychological explanation can be largely autonomous with respect to the mechanistic explanations of cognitive science, this fact does not pose a danger to a thoroughly naturalistic outlook on the nature of mind and cognition.