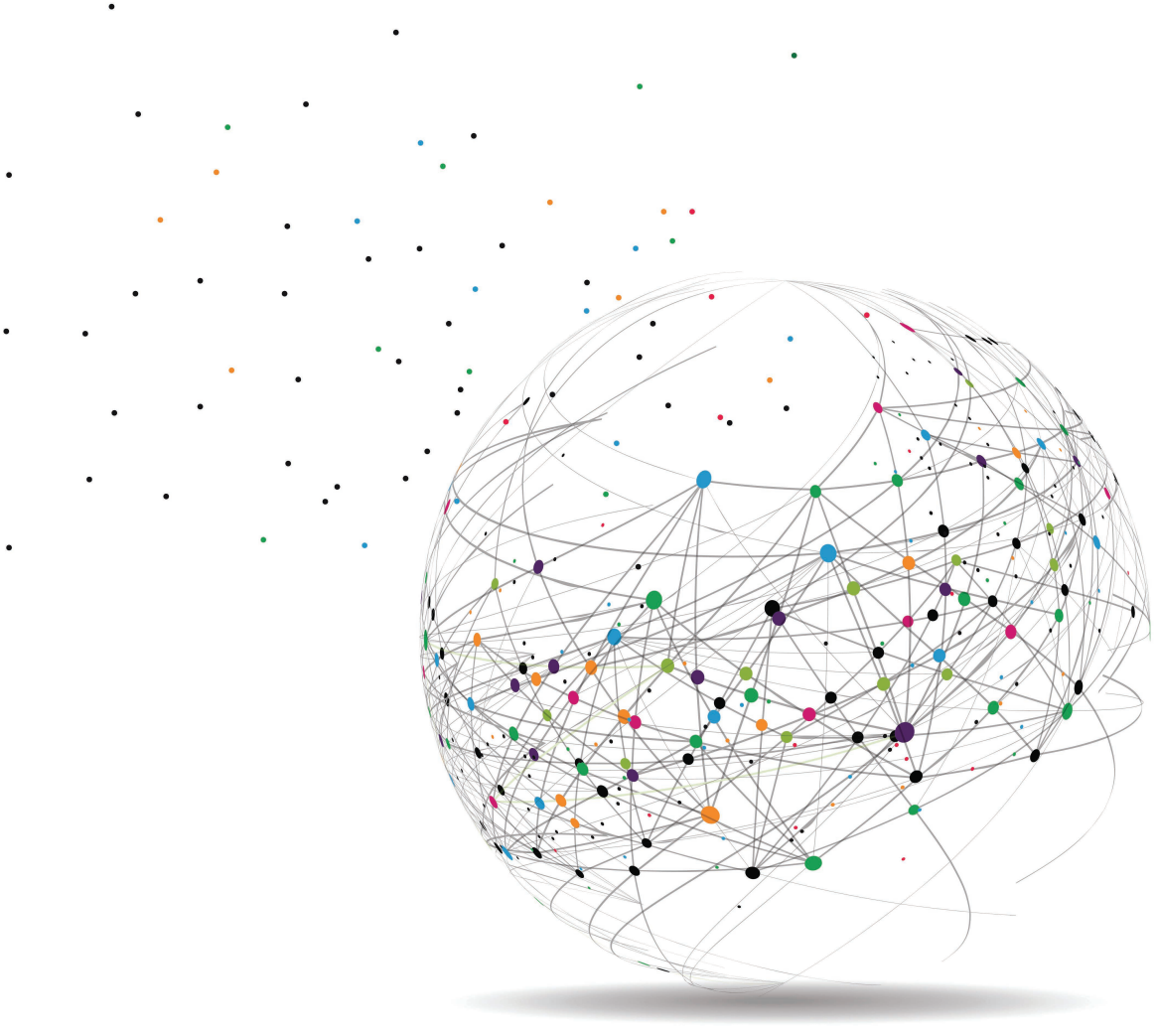
Michał Piekarski

Mechanizmy predykcyjne i ich normatywność

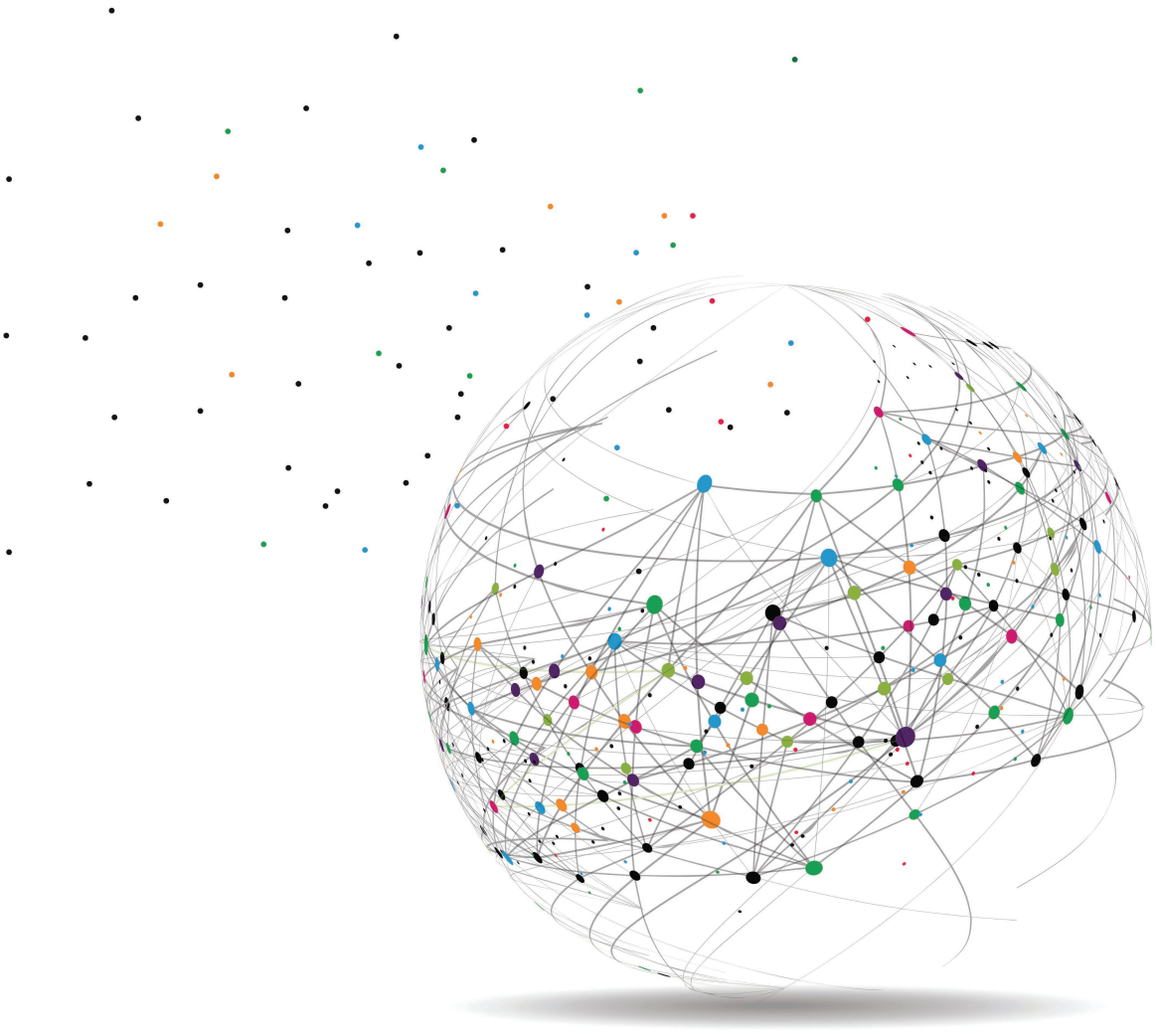


Michał Piekarski

Mechanizmy predykcyjne i ich normatywność

Michał Piekarski

Mechanizmy predykcyjne i ich normatywność



Michał Piekarski
Mechanizmy predykcyjne i ich normatywność

Recenzenci:
dr hab. Adriana Schetz, prof. US (Uniwersytet Szczeciński)
prof. dr hab. Paweł Kawalec (KUL JPII)

Projekt okładki:
Dominika Karaś

Ilustracja na okładce: shutterstock.com

Korekta językowa:
Elżbieta Matwiejczyk

Skład i łamanie:
Studio DTP Academicon | Patrycja Waleszczak
dtp@academicon.pl | dtp.academicon.pl

Publikacja jest udostępniona na licencji Creative Commons: Uznanie autorstwa 4.0 Polska.
Treść licencji jest dostępna na stronie: https://creativecommons.org/licenses/by/4.0/

Warszawa: Wydawnictwo Liberi Libri, 2020
www.LiberiLibri.pl

Wersja drukowana: ISBN 978-83-63487-44-7
DOI: 10.47943/lib.9788363487447

Spis treści

SPIS RYSUNKÓW I TABEL 9

WPROWADZENIE 11

PRZEDMOWA 29

ROZDZIAŁ I. POJĘCIE PRZETWARZANIA PREDYKCYJNEGO 35
Wstęp 35
1.1. Ogólna charakterystyka i podstawowe pojęcia 36
1.1.1. Predykcja 37
1.1.2. Model generatywny 38
1.1.3. Precyzja 41
1.1.4. Hierarchiczność 43
1.2. Możliwe algorytmy przetwarzania predykcyjnego i filozoficzne inspiracje 45
1.3. Kant i przetwarzanie predykcyjne 50
1.4. Krytyka przetwarzania predykcyjnego 57
1.4.1. Konserwatywne przetwarzanie predykcyjne 59
1.4.1.1. Internalizm a koce Markowa 61
1.4.1.2. Problem sceptyczny 63
1.4.1.3. Rzeczywiste wzorce w przetwarzaniu predykcyjnym 65
1.4.1.4. S-reprezentacje i realne wzorce 71
1.4.2. Pośredniość percepcji w przetwarzaniu predykcyjnym 74
1.5. Radykalne przetwarzanie predykcyjne 79
1.5.1. Psychologia ekologiczna 80
1.5.2. Enaktywizm 83
1.5.3. Psychologia ekologiczna, enaktywizm a przetwarzanie predykcyjne 86
1.5.4. Propozycja Nico Orlandi 87
Podsumowanie 90

ROZDZIAŁ II. PRZETWARZANIE PREDYKCYJNE JAKO BAYESOWSKI MODEL EKSPLANACYJNY 91
Wstęp 91
2.1. Uwagi wprowadzające 92
2.2. Podstawowe pojęcia 94
2.3. Teza mózgu bayesowskiego 103
2.4. Poziomy analizy a przetwarzanie predykcyjne 105
2.5. Problemy z modelami bayesowskimi 108
2.5.1. Problemy eksplanacyjne bayesizmu 110
2.5.2. Inżynieria odwrotna w przetwarzaniu predykcyjnym 113
2.6. Przetwarzanie predykcyjne jako bayesizm oświecony 115
2.6.1. Argumenty za eksplanacyjnym charakterem bayesizmu 115
2.6.2. Normatywność bayesowskiego modelu przetwarzania predykcyjnego 120
2.6.2.1. Normatywność a poziomy opisu 123
2.6.2.2. Preskryptywizm epistemiczny i ontyczny 127
Podsumowanie 128

ROZDZIAŁ III. ZASADA ENERGII SWOBODNEJ W PRZETWARZANIU PREDYKCYJNYM 131
Wstęp 131
3.1. Uwagi wprowadzające 132
3.2. Pojęcie zasady energii swobodnej 133
3.2.1. Pojęcie informacji Shannona 134

3.2.2. Homeostaza w zasadzie energii swobodnej 136
3.3. Koce Markowa a przetwarzanie predykcyjne 139
3.4. Energia swobodna a ciągłość życia i umysłu 142
3.4.1. Ujęcie enaktywistyczne 143
3.4.2. Krytyka ujęcia enaktywistycznego 147
3.4.3. Ujęcie kognitywistyczne: argumenty 151
3.4.4. Ujęcie kognitywistyczne: propozycja Hohwy'ego 158
3.4.5. Ujęcie nonkognitywistyczne a ujęcie kognitywistyczne 161
3.5. Trudności eksplanacyjne zasady energii swobodnej 164
3.6. Zastosowanie zasady energii swobodnej w przetwarzaniu predykcyjnym 173
Podsumowanie 175

ROZDZIAŁ IV. FUNKCJE I MECHANIZMY NORMATYWNE W KONTEKŚCIE PRZETWARZANIA PREDYKCYJNEGO 177
Wstęp 177
4.1. Uwagi wprowadzające 178
4.2. Normatywność homeostazy 180
4.3. Pojęcie funkcji 185
4.3.1. Funkcje w ujęciu Wrighta i Cumminsa 186
4.3.2. Funkcje w ujęciu Woutersa 190
4.3.3. Funkcje w ujęciu Millikan 194
4.3.3.1. Funkcje właściwe a mechanizmy predykcyjne 195
4.3.3.2. Krytyka podejścia teleosemantycznego 198
4.3.4. Funkcja reprezentowania 201
4.3.4.1. Funkcje w ujęciu Bickharda 203
4.3.4.2. Normatywność reprezentacji 208
4.3.4.3. Zarzuty Bickharda wobec przetwarzania predykcyjnego 209
4.4. Model normatywności dla przetwarzania predykcyjnego 211
4.5. Przetwarzanie predykcyjne a mechanistyczny model wyjaśniania naukowego 216
4.5.1. Normatywność mechanizmów predykcyjnych a preskryptywizm ontyczny 216
4.5.2. Mechanistyczny model wyjaśniania naukowego a mechanizmy normatywne 218
4.5.2.1. Mechanizmy predykcyjne jako przyczyny a subosobowy poziom wyjaśniania 220
4.5.2.2. Możliwość mechanistycznego wyjaśnienia normatywności 225
4.5.2.3. Wyjaśnienie mechanizmów normatywnych a ograniczenia 229
Podsumowanie 232

ROZDZIAŁ V. MECHANIZMY NORMATYWNE A DZIAŁANIA W PRZETWARZANIU PREDYKCYJNYM 233
Wstęp 233
5.1. Uwagi wprowadzające 234
5.2. Problem ciemnego pokoju 237
5.2.1. Odpowiedzi zwolenników przetwarzania predykcyjnego 239
5.2.2. Krytyka przetwarzania predykcyjnego z perspektywy problemu ciemnego pokoju 242
5.3. Działanie w warunkach niepewności 244
5.3.1. Wnioskowanie aktywne 246
5.3.2. Podejmowanie decyzji 248
5.3.2.1. Klasyczne modele 249
5.3.2.2. Krytyka klasycznych modeli z perspektywy koncepcji wnioskowania aktywnego 251
5.3.2.3. Rodzaje uczenia się a aktywne wnioskowanie 255
5.3.3. Uwagi wobec ramy aktywnego wnioskowania 258
5.4. Normatywne mechanizmy predykcyjne a ograniczenia 261
5.4.1. Mechanizmy normatywne: motywacja a nastawienie 262
5.4.2. Ograniczenia 268
5.4.2.1. Znaczenie ograniczeń dla działania 272

5.4.2.2. Ograniczenia w przetwarzaniu predykcyjnym 273
5.4.2.3. Afordancje – wstępna analiza 276
5.4.2.4. Granice wyjaśniania w przetwarzaniu predykcyjnym a postulat integracji 278
5.4.2.5. Ograniczenia jako warunki integracji modeli 279
5.5. Wyjaśnienie mechanizmów normatywnych w kontekście przetwarzania predykcyjnego 281
5.5.1. Problem ciemnego pokoju a mechanizmy normatywne 283
5.5.2. Rola predykcji 286
5.6. Mechanizmy normatywne a hipoteza rozproszonej normatywności 288
5.7. Predykcje skierowane na afordancje 291
5.8. Wnioskowanie aktywne jako wnioskowania enaktywne 295
Podsumowanie 298

ZAKOŃCZENIE 301

BIBLIOGRAFIA 309

PREDICTIVE MECHANISMS AND THEIR NORMATIVITY. SUMMARY 337

INDEKS NAZWISK 341

INDEKS RZECZY 349

TABLE OF CONTENTS 357

Spis rysunków i tabel

Rysunek 1. Schematyczne przedstawienie poziomów przetwarzania informacji przez model generatywny 44
Rysunek 2. Graficzne przedstawienie ramek zawierających skończone zbiory danych z określoną (procentowo) ilością szumu 66
Rysunek 3. Acykliczny graf skierowany sieci bayesowskiej 100
Rysunek 4. Acykliczny graf skierowany koca Markowa 102
Rysunek 5. Struktura przetwarzania komunikatu 135

Tabela 1. Porównanie konserwatywnego i radykalnego PP ze względu na ich inspiracje, teoretyczne zobowiązania i przyjęty model wyjaśniania 90

„…trying to understand perception by studying only neurons
is like trying to understand bird flight by studying only feathers.
It cannot be done"
(Marr, 1982, s. 27)

Wprowadzenie

Pojęcie normatywności jest jednym z najstarszych terminów filozoficznych. Jego historia sięga początków myślenia filozoficznego i wiąże się z odróżnieniem tego, jak się rzeczy mają, od tego, jak się mieć powinny. Dawni filozofowie wiązali je z moralnością i mówieniem o tym, co należy robić, a czego robić się nie powinno. Christine Korsgaard w swojej pracy poświęconej źródłom normatywności stwierdza, że kiedy poszukuje się filozoficznych podstaw moralności, to nie powinno nas interesować wyjaśnienie określonego rodzaju praktyk, lecz podanie odpowiedzi na pytanie dotyczące uzasadnienia przekonania, że moralność ucieleśniona w normach moralnych nie tylko w jakiś sposób oddziałuje na ludzi, lecz także, że powinna na nich oddziaływać. To pytanie Korsgaard nazywa „normatywnym". Na pytanie to mają składać się następujące kwestie: pierwsza z nich dotyczy znaczenia pojęć moralnych; druga sposobu ich użycia; zaś trzecia poświęcona jest pochodzeniu tych pojęć (Korsgaard, 1996, s. 10–11). Zdaniem badaczki te trzy pytania pozwalają sformułować teorię pojęć moralnych, która jednocześnie ma być teorią normatywności.

Propozycja Korsgaard, aby myśleć o normatywności za pomocą pojęć moralnych, zdradza, czy wręcz wyraża, pewną ogólną postawę bądź przekonanie dotyczące tego, czym normatywność jest i kto się nią zajmuje. Normatywność w tej perspektywie jest problemem badanym przez etyków i zasadniczo dotyczy zagadnień takich jak: obowiązywalność zasad czy reguł; kwestia powinności i obowiązków; błąd naturalistyczny; problem naturalizacji etyki. Problem normatywności, pomimo że powiązany z filozofią przynajmniej od czasów Platona, zostaje w jakimś stopniu wyrażony dopiero w filozofii nowożytnej oraz związany z etyką i prawem, potem zaś z metodologią, logiką czy epistemologią (por. Brożek, 2013, s. 19–44; Duchliński, 2015, s. 16–62). Należy dodać, że kiedy używa się pojęcia normatywności, to chce się powiedzieć coś o rzeczywistości, co w jakimś sensie wykracza poza deterministyczny porządek przyczynowo-skutkowy. Od razu pojawia się jednak zarzut sformułowany przez Davida Hume'a, który głosi, że nie ma przejścia od zdań o faktach do zdań o powinnościach. Jego pogląd do dziś jest obecny w dyskusjach filozoficznych i wielu autorów twierdzi, że w naturalistycznie

zorientowanej filozofii, ale również nauce, nie ma miejsca na normatywność[1]. Takie stwierdzenie budzi jednak od razu pewne zastrzeżenie (por. Bickhard, 2014). Nie wiadomo bowiem, co właściwie należy rozumieć przez normatywność. Wydaje się, że sam Hume zajmował się normatywnością w kontekście moralności i prawa. Również dzisiaj jest wielu autorów, którzy – tak jak Korsgaard – mówiąc o tym, co normatywne, mają ostatecznie na myśli to, co moralne.

Utożsamienie moralności z normatywnością budzi jednak duże zastrzeżenia, ponieważ jest to wyraźnie zbyt wąskie ujęcie normatywności. Wystarczy, przejrzeć opracowania poświęcone temu zjawisku bądź dokonać szybkiego przeglądu baz bibliograficznych, zawierających jako słowo kluczowe termin „normatywność", aby spostrzec, że nie da się mówić o normatywności tylko w jednym sensie czy z jednej perspektywy. Badacze piszą bowiem m.in. o normatywności znaczenia, percepcji czy przekonań, ale także o normatywności logiki, prawa czy matematyki. Pisze się o normatywnych wymogach, procesach, funkcjach czy modelach. W perspektywie obecnie toczonych dyskusji moralność wraz z całą sferą etyczną należy raczej uznać za pewną węższą klasę normatywności niżeli za jej odpowiednik bądź zamiennik. Z tego też powodu badania nad normatywnością mogą zupełnie abstrahować od kwestii moralnych czy praktycznych.

Interesującą typologię zaproponował Roman Ingarden. Co prawda odnosi się ona do pojęcia normy, ale jak sądzę, może zostać z powodzeniem zastosowana w odniesieniu do pojęcia normatywności. Ingarden odróżnił: (1) znaczenie przedmiotowe (obowiązujący wzór, schemat; to, co normalne, w przeciwieństwie do anormalnego itd.) – normatywne są zatem schematy i wzorce[2]; (2) znaczenie w sensie bycia stanem rzeczy (to, co powinno zajść; to, co ma być wykonane; ustanowiony powinny stan rzeczy, np. waga monet w danym kraju); (3) znaczenie logiczne (sąd o powinnym stanie rzeczy; myśl ustanawiająca, jakie coś powinno być; rozkaz, że coś ma być takie a takie); (4) znaczenie językowe (zdanie wyrażające pewne normy); (5) znaczenie poznawcze (kryterium poznawcze, np. prawda) (Ingarden, 1989, s. 17, 143–144).

Do typologii zaproponowanej przez Ingardena dodałbym jeszcze jedno znaczenie, w jakim można mówić o normatywności i którym będę się w tej pracy posługiwał: znaczenie konstytutywne – normatywne jest to, co konstytuuje dany przedmiot,

[1] Jest to pogląd podzielany nie tylko przez filozofów, ale również przez wielu psychologów i kognitywistów (por. podrozdział *Problemy z modelami bayesowskimi* oraz podrozdział *Normatywność bayesowskiego modelu przetwarzania predykcyjnego*).

[2] Innymi słowy, normatywne jest to, co ma charakter regularny, powtarzalny, nielosowy. Takimi normatywnymi własnościami są np. stała wielkość i kształt postrzeganego przedmiotu (tzw. stałości percepcyjne), które normują procesy percepcyjne. „Widzimy, że stała wielkość i kształt oddziałują jako norma, a jej uchwycenie można nazwać uprzywilejowanym, ponieważ zapewnia osiągnięcie celu procesu spostrzegania – jasną i wyraźną artykulację pola" (Maciejczak, 2005, s. 191).

stan rzeczy bądź proces. Jest normatywne, ponieważ określa warunki, jakie *musi* spełnić dany przedmiot, stan rzeczy albo proces, aby można było o nim orzekać, że jest właśnie tym, a nie innym. Przez konstytucję rozumiem tutaj proces, który umożliwia pojawienie się danego przedmiotu, stanu rzeczy bądź procesu. Owo pojawienie trzeba rozumieć nie w sensie pewnej możliwości fizycznej, biologicznej itd., ale w sensie logicznym[3]. W niniejszej pracy twierdzę bowiem, że mechanizm jest normatywny, kiedy konstytuuje określone działania i zachowania[4]. Jest on z jednej strony ich przyczyną, z drugiej zaś – jak wykażę – wyznacza pewne pole możliwości, spośród którego podmiot, *resp.* organizm, kierowany własnymi celami, preferencjami, uwarunkowaniami środowiskowymi itd.[5], wybiera odpowiednie w danej sytuacji i wedle własnego rozeznania działanie. Funkcję konstytutywną takiego mechanizmu będę wiązał z generowanymi przezeń predykcjami, które są konstytutywne dla działań w tym sensie, że określają, jakie działania są uzasadnione, a jakie nie, biorąc pod uwagę określoną sytuację itd. Innymi słowy, predykcje wskazują na warunki, jakie powinno spełniać działanie, aby sprostać określonym wymogom biologicznym, poznawczym, społecznym itd. Będę twierdził, że pomiędzy predykcją a podjętym działaniem zachodzi tzw. relacja motywacyjna, bez istnienia której nie da się wyjaśnić faktu funkcji kierowania działaniami przez podmiot. Oznacza to, że wyjaśnienie tej funkcji z konieczności musi odwołać się do mechanizmu predykcyjnego je konstytuującego. Z tego też powodu będę pisał o normatywności mechanizmów jako eksplanacyjnej albo konstytutywnej, wiążąc bezpośrednio ze sobą oba te sensy[6].

---

[3] Wiąże się to z tzw. warunkami logicznymi, które określają kryteria wyboru danego działania.

[4] Przez zachowania rozumiem (niskopoziomowe, np. na poziomie fizjologii oka) skoordynowane postępowanie danego organizmu w określonym środowisku. Przez działania zaś rozumiem takie (wysokopoziomowe) zachowania, jakim w języku psychologii potocznej jesteśmy skłonni przypisać cele, intencje i zamiary (pojęcie poziomów zostanie wyjaśnione w dalszych partiach niniejszej książki). Przykładowo: odruch rogówkowy będzie zachowaniem, zaś próba zatrzymania tego odruchu w celu włożenia szkła kontaktowego będzie już działaniem. W niniejszych analizach interesują mnie głównie poznawcze, behawioralne i biologiczne mechanizmy, dzięki którym takie działania mogą być realizowane. Analiza działań świadomych, ich intencjonalności, celowości znacznie wykracza poza ramy przeznaczone dla tych rozważań. Jednak zaproponowana przeze mnie perspektywa z pewnością może być zastosowana do tego typu problemów przy założeniu, że zostanie ona wzbogacona przez odpowiednie modele czy rozwiązania filozoficzne.

[5] Czynniki te określają kryteria wyboru danego działania.

[6] Zaproponowane przeze mnie pojęcie konstytucji jest w pewnym stopniu inspirowane pracami fenomenologów (por. Husserl, 1974, s. 1982), zrywa ono jednak zupełnie z transcendentalnym wymiarem, który jest charakterystyczny dla ujęcia fenomenologicznego. Na temat transcendentalizmu w odniesieniu do PP por. podrozdział *Kant i przetwarzanie predykcyjne*.

Broniona w tej pracy koncepcja konstytucji ma sens naturalistyczny, bowiem nie odnosi się do warunków transcendentalnych, ale warunków, które określam jako logiczne. Twierdzę bowiem, że nie można w badaniach różnych zjawisk przyjąć jakiejś zewnętrznej albo uprzywilejowanej perspektywy transcendentalnej, ale jedynie względną, zrelatywizowaną do określonych warunków biologiczno-poznawczych perspektywę badawczą. Trafnie prezentowane tutaj nastawienie opisuje Leszczyński: „Badamy warunki możliwości wszelkiego ludzkiego poznania. Jednak nie badamy ich od razu jako pewnej całości, zamkniętego i skończonego systemu, ale raczej wychodząc od poszczególnych procesów, na przykład percepcji czasu i przestrzeni, sposobu indywiduacji i organizacji przedmiotów w ramach pola wzrokowego, procedur idealizacji itd. Przy czym tych ostatnich procesów również nie badamy całościowo, lecz wychodząc od konkretnych i partykularnych sytuacji poznawczych. Oczywiście, w tym poznaniu cały czas biorą udział owe ogólne warunki, ku których rekonstrukcji zmierzamy,

Współcześnie pojęcia, procesy oraz mechanizmy normatywne są przedmiotem badanym z jednej strony przez filozofię języka, filozofię umysłu czy filozofię biologii, z drugiej zaś przez kognitywistykę i niektóre nauki społeczne. Wydaje się jednak, że istnieje tendencja do odchodzenia od myślenia o normatywności tylko w kategoriach moralno-prawnych na rzecz myślenia o niej jako zjawisku naturalnym, którego pochodzenie da się wyjaśnić metodami naukowymi.

Na wstępie należy jednak stwierdzić, że nie ma jednoznacznej zgody w kwestii rozumienia tego, czym jest normatywność. Wydaje się bowiem, że z jednej strony spór o nią jest sporem ontologicznym, który dotyczy tego, jakiego rodzaju bytami lub przedmiotami są normy, wartości czy same powinności, z drugiej zaś, śledząc współczesną dyskusję nad normatywnością, można odnieść wrażenie, że spór o nią jest metaprzedmiotowym sporem dotyczącym pojęć, których się używa, sposobów uzasadniania norm czy problemem ich ewentualnej naturalizacji. Autorzy w różny sposób definiują normatywność, często wydaje się, że arbitralnie, nie zważając na to, co głoszą ich oponenci.

Można wskazać następujące rozumienia pojęcia normatywności we współczesnych ujęciach, nie określając jednocześnie jakiegoś wybranego kryterium ich doboru (por. Bicchieri, 2006; Duchliński, 2015, s. 28–33; Henderson, 2017, s. 409–435)[7]:

1. Normatywność wiąże się z podmiotową umiejętnością (kompetencja normatywna) do tworzenia różnego rodzaju norm, reguł, zasad, które określają ludzką praktykę (por. m.in. Waleszczyński, 2018).

2. Normatywność to własność określonych standardów etycznych, czyli sądów odnoszących się do tego, co być powinno, i tego, co być nie powinno (por. m.in. Korsgaard, 1996).

3. Normatywność to własność takich wypowiedzi bądź wyrażeń językowych, jak zdania powinnościowe, imperatywy, pytania, twierdzenia, normy itd. (por. m.in. Kripke, 1982; Boghossian, 1989, 2003; Whiting, 2013).

4. Normatywność to własność reguł rozumianych jako obiektywne racje działania, odróżnialne od motywów czy bodźców (por. m.in. Anscombe, 1957; Davidson, 1963; Brożek, 2013; Piekarski, 2016; Henderson 2017, s. 410).

5. Normatywność to pewnego szczególnego rodzaju własność rzeczywistości, pewnych przedmiotów, stanów rzeczy bądź zdarzeń (por. m.in. Hartman, 1926; Ingarden, 1989; Moń, 2011; Scheler, 2014).

6. Normatywność to własność etycznych i prawnych systemów normatywnych, tj. praw, kodeksów itd. Należy ją wiązać z określonymi zbiorami zdań praktycznych

jednak my za każdym razem ujmujemy je od innej, szczegółowej formy, starając się następnie 'skleić' nasze konkluzje i wnioski w pewne ogólne twierdzenia" (Leszczyński, 2010, s. 149).

Należy również dodać, że w mechanicyzmie, który stanowi istotne metodologiczne odniesienie dla prezentowanych tutaj analiz, mówi się o wyjaśnieniach konstytutywnych. Takie wyjaśnienia tłumaczą, w jaki sposób rzeczy mają potencjał przyczynowy, powołując się na ich części oraz sposób organizacji (Salmon, 1984; por. także Ylikoski, 2013). Mechanizmy konstytutywne zaś, w moim rozumieniu, wiążą się z takimi wyjaśnieniami, niemniej pozwalają one również wskazać określone warunki logiczne dla generowanych przezeń działań.

[7] Lista ta z całą pewnością nie wyczerpuje wszystkich możliwości.

i teoretycznych, powiązanych związkami logicznymi i pozalogicznymi (por. m.in. Kelsen, 1945; Kalinowski, 1960).

7. Normatywność to własność epistemiczna pojęć uzasadniających wypowiedzi, sądy itd. Wiąże się bezpośrednio z takimi pojęciami jak: racjonalność, rozum czy racja (por. m.in. Brandom, 1994, 2000; Broome, 2007).

8. Normatywność to własność interakcji społecznych. Normatywne są oczekiwane sposoby zachowania się jednostki w danej grupie społecznej, pewne wzorce zachowań czy zasady regulujące sposoby działania (por. m.in. Bicchieri, Muldoon, Sontuoso, 2014; Henderson, 2017, s. 409; Wachowski, 2018a).

9. Normatywność to psychologiczna własność podzielana przez daną populację. Swoiste wyuczone zachowanie odniesione do określonych indywidualnie warunków (por. m.in. Darden, 2006).

10. Normatywność to własność homeostatyczna, polegająca na utrzymywaniu stałości parametrów danego systemu w odniesieniu do środowiska, w jakim on funkcjonuje (por. m.in. Friston, Stephan, 2007; Bickhard, 2014).

11. Normatywność to własność tzw. funkcji normatywnych, pozwalająca na wyjaśnienie możliwości dysfunkcji (por. m.in. Millikan, 1989, 2011; Christensen, Bickhard, 2002; Bickhard, 2014).

12. Normatywność w swoim podstawowym wymiarze odnosi się do własności najprostszych zachowań organizmów niezłożonych, np. bakterii (por. m.in. Barandiaran, Di Paolo, Rohde, 2009; Barandiaran, Egbert, 2011, 2013).

Przegląd zaproponowany przeze mnie, podobnie jak wspomniana typologia Ingardena, nie wyczerpuje jednak wielości odniesień do normatywności w filozofii współczesnej. Wydaje się także, że te różnorodne sposoby rozumienia czy postrzegania normatywności oparte są na bardziej podstawowym wymiarze dociekań, który należy wiązać z dyskusją toczoną przez zwolenników naturalizmu z ich przeciwnikami. Opozycja naturalizm – antynaturalizm nie jest w filozofii nowa, ale w kontekście sporu o normatywność zyskuje nowe znaczenie.

Naturalizm, najogólniej rzecz ujmując, opiera się na następujących twierdzeniach:

1. Normatywność można sprowadzić do moralności (por. Davies, 2001; Copp, 2009; Churchland, 2013; de Waal, 2013).

2. Moralność nadbudowana jest na zdolnościach społecznych, które powstały w czasie ewolucji biologicznej. Co ważne, są one wspólne także dla innych naczelnych (Haidt, 2001; Tomasello, 2015; de Waal, 2013, 2014).

3. Głównym zadaniem moralności jest łagodzenie konfliktów w grupie. W tym znaczeniu moralność jest przede wszystkim pewnego rodzaju mechanizmem społecznym (por. de Waal, 2013, 2014).

4. Moralność nie jest wyrazem racjonalności, ale głównie jest ugruntowana w afektach i przeżywanych emocjach (por. Haidt, 2001; Hume, 2007).

5. Systemy normatywne są efektem ewolucji kulturowej (por. Tomasello, 2015).

Charakterystyczne dla naturalizmu jest przekonanie, że badanie pracy oraz struktury mózgu może prowadzić do uzyskania istotnych informacji na temat moralności i normatywności. W mechanizmach neuronalnych, a konkretniej w wyjaśnieniach ich działania, ma tkwić odpowiedź na pytanie o normatywność. Nie oznacza to, że wyjaśnienie pracy mózgu pozwoli rozwiązać wszelkie problemy i wyeksplikować pojęcia normatywne. Znaczy to tylko tyle, że podejście budowane na podstawie takich wyjaśnień pozwoli dać taką odpowiedź na pytanie o źródła normatywności, jaka nie będzie musiała odwoływać się do jakichś nadnaturalnych własności czy bytów albo pojęć racjonalności bądź powinności[8].

Zwolennicy antynaturalizmu są grupą równie niejednorodną jak naturaliści. To, co ich łączy, to podzielanie dwóch podstawowych przekonań, które mogą być na różne sposoby formułowane i akcentowane:

1. Twierdzeń normatywnych nie można w pełni wyeksplikować przy użyciu deskryptywnych pojęć języka nauk przyrodniczych[9].
2. Własności normatywne nie są sprowadzalne do własności naturalnych i należy je od nich odróżnić (por. Audi, 2009, s. 92)[10].

Badacze (por. Ridge, 2003) zwracają uwagę na podstawową trudność związaną z charakterystyką antynaturalizmu, dotyczącą wspomnianego odróżnienia własności naturalnych od nienaturalnych (normatywnych). Własności naturalne są charakteryzowane m.in. jako takie własności, jakie (1) są przedmiotem nauk przyrodniczych (Moore, 1903, s. 40); (2) są wykorzystywane w wyjaśnieniach naukowych (Little, 1994); (3) mogą zostać zidentyfikowane przez najlepszą teorię naukową i można je opisać pojęciami dostępnymi dla istoty zajmującej nielokalny punkt widzenia na świat (Crisp, 1996); (4) można je poznać tylko *a posteriori* (Moore, 1903, s. 39; Copp, 2007); (5) mogą istnieć same w czasie (Moore, 1903, s. 41); (6) mają moce przyczynowe (Lewis, 1983). Najogólniej mówiąc, etyczny czy normatywny antynaturalizm można scharakteryzować jako pogląd, zgodnie z którym istnieją prawdziwe etyczne własności i fakty, które nie należą do naturalnych własności i faktów świata. Antynaturalizm jest zatem etycznym realizmem, który odrzuca naturalistyczne konstrukty etycznych własności i faktów (por. FitzPatrick, 2011).

Po przedstawieniu tej panoramy poglądów trzeba wskazać, jakie miejsce w niej zajmuje moje stanowisko badawcze. Przede wszystkim należy podkreślić, że per-

[8] W tym kontekście warto przytoczyć słowa Patrycji Churchland: „Według mojej hipotezy dotyczącej neuronalnych podstaw postępowania moralnego, moralność jest tak prawdziwa, jak prawdziwe są zachowania społeczne. Prawdziwie ludzkie zachowania moralne, w całej swej okazałości i złożoności, nie mogą zostać zdewaluowane przez fałszywy dylemat: albo Bóg jest gwarantem prawa moralnego, albo moralność jest iluzją. Jest to fałszywy dylemat, ponieważ moralność może być – i twierdzę, że *jest* – zakorzeniona w naszej biologii, w naszej zdolności do współczucia i naszej zdolności do uczenia się i rozumienia" (Churchland, 2013, s. 320).

[9] Zdaniem Roberta Audiego (2013, s. 38) negacja tego przekonania jest kluczowa dla współczesnego naturalizmu.

[10] Naturalizm ten, na co należy zwrócić uwagę, może mieć m.in. oblicze redukcjonistyczne, kiedy twierdzi się, że własności normatywne da się sprowadzić do naturalnych, lub nieredukcjonistyczne, kiedy twierdzi się, że np. własności normatywne superweniują na własnościach naturalnych.

spektywę przyjętą przeze mnie w tej pracy można określić jako naturalistyczną, ponieważ za punkt wyjścia przyjmuje ona dane dostarczane m.in. przez kognitywistykę, neuronaukę, psychologię poznawczą czy biologię, a więc tzw. nauki szczegółowe. Nie podzielam jednak przekonania wielu zwolenników naturalizmu (m.in. redukcjonistów), według których wyniki nauk szczegółowych, np. neuronauki, mogą dostarczyć nam ostatecznej odpowiedzi na filozoficzne pytania dotyczące natury moralności, normatywności, powinności czy słuszności. Twierdzę bowiem, że nauki te mogą w pewnym stopniu, dostarczając odpowiednich danych oraz rozwiązań, pomóc nam odpowiedzieć na pytanie, dlaczego człowiek jest istotą „podatną" na normy, reguły czy pewne wzorce. Innymi słowy, interesuje mnie odpowiedź na pytanie o to, co czyni nas jako ludzi „istotami normatywnymi", i odpowiedzi tej poszukuję wśród danych, jakie dostarczają nam badania z zakresu kognitywistyki, biologii czy neuronauki. W tym sensie moje badania można określić jako należące do filozofii kognitywistyki, rozumianej jako jeden z rodzajów filozofii nauki.

W niniejszej pracy będę bronił przekonania, zgodnie z którym procesy poznawcze zawierają w sobie pewien normatywny komponent obecny w sposobie, w jaki doświadczamy świata i innych ludzi. Tym komponentem są określone mechanizmy poznawcze, nazywane przeze mnie „normatywnymi". Są one normatywne, ponieważ są konstytutywne dla niektórych działań i zachowań. Ich normatywność wywodzi się z niektórych funkcji, jakie mogą spełniać[11]. W tej pracy zaproponuję analizę mechanizmów predykcyjnych, których podstawową funkcją jest generowanie predykcji. Predykcje, jak wykażę, mają charakter normatywny ze względu na ich konstytutywną rolę w kierowaniu działaniami. Przez predykcję najogólniej rozumiem statystyczne przewidywanie informacji pochodzących z sygnału zmysłowego (por. podrozdział *Predykcja*). Mechanizm generujący tak rozumiane predykcje określam mianem „mechanizmu normatywnego". Pojęcie to zostanie w pełni wyeksplikowane w dalszych partiach pracy.

Ze względu na przedmiot tej książki oraz jej założenia należy powiedzieć, że przyjmuję tutaj stanowisko naturalizmu metodologicznego, zgodnie z którym, najogólniej stwierdzając, filozofia i nauka są zaangażowane zasadniczo w to samo przedsięwzięcie badawcze, dążą do podobnych celów i wykorzystują podobne metody badawcze (Papineau, 2007). Naturalizmu w tym sensie nie traktuję jako doktryny, ale raczej jako postulat metodologiczny, według którego należy szukać naturalnych wyjaśnień dla badanych przez nas aspektów świata (Giere, 2001). W praktyce oznacza to porzucenie postulatów ugruntowania wiedzy, np. w transcendentalizmie. Nie przyjmuję jednak naturalizmu ontologicznego, który zakłada, że wszystkie byty muszą być identyczne lub metafizycznie redukowalne do bytów fizycznych [tj. niementalnych i niewitalnych (Montero, Papineau, 2005, s. 233–237)], naturalizmu epistemologicznego, którego większość zwolenników odrzuca normatywne

[11] W pewnym sensie bronię zatem specyficznej wersji realizmu normatywnego.

roszczenie epistemologii jako dyscypliny filozoficznej, zastępując ją psychologią poznawczą albo neuronauką, oraz naturalizmu semantycznego, sugerującego redukcję wszelkich sensownych wypowiedzi o świecie do języka, którejś z nauk szczegółowych. Ujęcia te bowiem zakładają fizykalizm i redukcjonizm, tj. stanowiska, których przesłanek w niniejszych analizach nie zakładam. W praktyce oznacza to, że zajmuję się konstytuowaną biologicznie normatywnością mechanizmów poznawczych, niemniej nie odrzucam w żadnym stopniu zjawisk, własności czy przedmiotów normatywnych, które są konstytuowane w inny niż biologiczny czy fizyczny sposób. Proponowane przeze mnie ujęcie nie zakłada zatem redukcji pojęć normatywnych czy moralnych do pojęć fizykalnych czy biologicznych[12]. Bronię bowiem przekonania, że na poziomie biologicznym istnieją określone mechanizmy normatywne, dla których wyjaśnienia konieczne jest wykorzystanie pojęć i narzędzi zaczerpniętych z nauk przyrodniczych. Odrzucam jednak przekonanie, zgodnie z którym normatywność jako przedmiot badany przez filozofów można w pełni wyjaśnić jedynie za pomocą pojęć i narzędzi tych nauk. Analizy przeprowadzone w niniejszej pracy dotyczą zatem szczególnego pola badawczego i mają prowadzić do wypracowania określonego stanowiska teoretycznego na gruncie filozofii kognitywistyki i filozofii biologii. Nie jest zatem moim celem zredukowanie tradycyjnej filozoficznej problematyki normatywności do problemów biologii czy neuronauki, ponieważ pytania, które formułuję, i odpowiedzi, które staram się na nie podać, mają charakter ściśle epistemologiczny[13].

Po tych wstępnych i koniecznych uwagach mogę przejść do dalszych rozważań przedmiotowych.

W standardowym ujęciu naturalizm w filozofii zakładał porzucenie normatywnych zagadnień na rzecz ujęć deskryptywnych i opisów przyczynowych (por. Firth, 1964; Alston, 1976; Quine, 1986)[14]. Wielu badaczy podkreśla jednak konieczność wykorzystania pojęć normatywnych w wyjaśnieniach odpowiednich funkcji poznawczych i biologicznych (por. m.in. Bickhard, 2003, 2014; Christensen, 2012; Kitcher, 1993; Millikan, 2011). Z jednej bowiem strony twierdzą oni, że funkcje są normatywne, a ich normatywność pozwala wyjaśnić funkcjonowanie organizmów, jednocześnie pokazując, kiedy jest ono niepoprawne; z drugiej zaś atrakcyjność ich ujęcia polega na tym, że sprowadzają oni to, co normatywne, do tego, co opisowe. Wayne Christensen (2012) podkreśla konieczność uwzględnienia czegoś, co określa mianem normatywności oceniającej (*evaluative normativity*) w naukowych analizach

[12] W niniejszej pracy twierdzę, że nawet w obrębie wyjaśniania mechanizmów biologicznych redukcja do niższych poziomów mechanizmów jest zależna jedynie od względów pragmatycznych (np. cele albo strategia badawcza), a nie ontycznych. W praktyce oznacza to, że np. mechanizmu odpowiedzialnego za rozpoznawanie emocji drugiego człowieka nie trzeba redukować do określonych mechanizmów fizykochemicznych. Z tego też powodu mechanistyczny model wyjaśniania, który wykorzystuję w tej pracy, nie jest z zasady redukcjonistyczny.

[13] Epistemologia jest ukierunkowana bowiem na normatywną stronę poznania, podkreślając w ten sposób znaczenie uzasadnienia i poszukiwania racji.

[14] Z poglądem tym dyskutuje m.in. Jaegwon Kim (1988).

systemów żywych i sztucznych. Musimy bowiem być w stanie powiedzieć, kiedy takie, a nie inne stany danego systemu można określić jako lepsze bądź gorsze. Teza Christensena znajduje swoje potwierdzenie we wpływowej opinii Paula Thagarda, który stwierdził, że filozofia ma do odegrania niebanalną rolę, dostarcza bowiem nauce, a w szczególności kognitywistyce, określone pytania normatywne. Thagard (2009, s. 238) stwierdza, że:

> filozofia nie dotyczy tylko tego, jak się rzeczy mają, ale także tego, jakie powinny być. Filozoficzne teorie wiedzy i moralności muszą wykraczać poza opisowe teorie tego, jak ludzie myślą i działają, ale rozwijać także normatywne (nakazowe) teorie, w jaki sposób ludzie powinni myśleć i działać.

W tym sensie filozofia spełnia istotną funkcję wspomagającą naukę.

Pogląd Thagarda i Christensena z jednej strony wskazuje na ważne miejsce dla języka normatywnego w debatach naukowych, z drugiej zaś wyraża niewypowiedzianą wprost myśl, że normatywność jest czymś, co (1) ma charakter oceniający oraz że (2) jest ona przypisywana przedmiotom, stanom rzeczy, procesom czy funkcjom w sposób zewnętrzny. Czy zatem nie można o czymś orzekać, że jest normatywne, ponieważ takie jest? I czym miałaby być normatywność przypisana danemu przedmiotowi, jego własności bądź funkcji bez odniesienia do zewnętrznego obserwatora i bez potrzeby oceniania tego przedmiotu itd. jako poprawnego, lepszego czy racjonalniejszego?

Aby zilustrować przywołaną tutaj trudność, a także ułatwić czytelnikowi zrozumienie specyfiki proponowanego przeze mnie ujęcia normatywności, omówię pewien prosty przykład znany wszystkim z codziennego życia. Wyobraźmy sobie, że jestem umówiony na ważne spotkanie w sprawie pracy na drugim końcu miasta. Nie dysponuję samochodem, a najbliższy przystanek autobusowy znajduje się 20 minut drogi od mojego domu. Istotny jest zatem dla mnie np. stan pogody, ponieważ wybierając swój ubiór, a w dalszej perspektywie być może również środek transportu, muszę ten stan wziąć pod uwagę. Gdyby np. miało padać, to będę musiał założyć palto i zabrać parasol, gdyby zaś pogoda była słoneczna, to mogę zrezygnować z tych elementów ubioru. Patrzę zatem przez okno i widzę zachmurzone niebo. Mogę również spojrzeć na ciśnieniomierz oraz termometr. Moje wcześniejsze doświadczenia ze zmienną pogodą oraz informacje, których dostarczają mi zmysły (np. poziom zachmurzenia, wysokość ciśnienia atmosferycznego, temperatura powietrza), pozwalają mi przewidzieć, że niebawem spadnie deszcz, a być może pojawi się burza. Trafność mojego przewidywania jest kluczowa dla podejmowanych przeze mnie działań (wyboru ubrania, środka transportu, czasu opuszczenia mieszkania itd.). Wiem bowiem, że załamanie się pogody może utrudnić moje dotarcie na spotkanie oraz dodatkowo zniszczyć ubiór. Moje przewidywanie dotyczące przyszłego stanu pogody warunkuje wybór takich, a nie

innych działań[15]. Oznacza to, że treść mojego przewidywania staje się normatywna$_1$ dla wyboru określonych działań (w sensie ich przyczyny konstytutywnej, tj. takiej, jaka określa warunki logiczne pojawienia się i realizacji danego działania)[16]. Nie stanowi ono oczywiście jakiejś normy działania w sensie etycznym bądź prawnym, ale jest tym czynnikiem, który powinienem wziąć pod uwagę, podejmując takie, a nie inne decyzje dotyczące w tym przypadku wyboru ubioru, a także tego, czy mam zamówić taksówkę, czy może odwołać spotkanie. Przywołany przykład dobrze pokazuje, że moje przewidywanie wpływa istotnie na wybór działań. Innymi słowy, tworząc takie, a nie inne przewidywanie, jestem zobowiązany podjąć takie, a nie inne działania[17]. W tym sensie to przewidywanie jest normatywne$_2$ dla działań, które ono warunkuje. Określa bowiem, które działania powinienem podjąć, jeżeli zakładam prawdziwość tego przewidywania. Wskazuje również te działania, które zupełnie nie wiążą się z owym przewidywaniem. Przewidując, że będzie burza, i chcąc zminimalizować jej negatywne skutki dla realizacji celu, jakim jest dotarcie na spotkanie o umówionej godzinie w schludnym ubiorze, mogę założyć palto, zamówić taksówkę bądź przełożyć spotkanie na inny termin. Nie zminimalizuję jednak tych negatywnych skutków, jeżeli włączę telewizor, zamówię pizzę albo pójdę spać, ponieważ nie zrealizuję podjętego zamiaru. Normatywność mojego przewidywania nie tylko wiąże się z tym, że określa ono to, co mogę zrobić, ale również z tym, że wyklucza wiele działań, które nie są istotne dla mojego przewidywania, jeżeli uznaje je za trafne bądź prawdziwe. Jest ona również normatywna$_3$ w tym sensie, że jeżeli będę działał zgodnie z moim przewidywaniem (tj. jego treścią),

[15] W typowym ujęciu teorii decyzji to tylko jeden z elementów, obok funkcji użyteczności oraz kryterium decyzyjnego. Jak jednak wykażę w rozdziale *Mechanizmy normatywne a działania w przetwarzaniu predykcyjnym*, elementy te są w pewnym stopniu sprowadzalne do innych wymogów.

[16] Pojęcie normatywności jest tutaj rozumiane warunkowo, a nie jako pojęcie bezwarunkowe, które określa taką, a nie inną normę. Jednak, za czym będę argumentował w podrozdziale *Rola predykcji*, statystyczna skuteczność działań, opartych na takich, a nie innych predykcjach i podejmowanych w takich, a nie innych warunkach, może prowadzić do wytworzenia określonych wzorców zachowań, które można wiązać z normatywnością rozumianą bezwarunkowo.

[17] Należy podkreślić, że zgodnie z matematyczną teorią gier w większości wypadków wyboru działań, ale w zależności od rozkładu użyteczności (ryzyka) może być kilka wyborów, jak pokazują różne dylematy związane ze stanem równowagi Nasha. Równowaga Nasha to takie pary strategii działania, które są najlepszymi (tj. optymalnymi) odpowiedziami na siebie nawzajem. Gdy w grze zostanie osiągnięta tak rozumiana równowaga, żaden z graczy nie może poprawić swojego wyniku poprzez jednostronną zmianę wybranej strategii. Z tak rozumianą równowagą wiążą się różne paradoksy (np. dylemat więźnia). W takim ujęciu zwrot „jestem zobowiązany" jest silnie uwarunkowany omawianym stanem równowagi, ale także jego kryterium, jakie zostało przyjęte.

Uwaga ta jednak, co tylko zasygnalizuję, może budzić wątpliwości, jeżeli chcemy ją wiązać z PP. Po pierwsze, niektórzy badacze wykazują, że ludzie np. w przeciwieństwie do szympansów, mają wyraźną skłonność do odrzucania ofert, gdy podział odchodzi od równości. Oznacza to, że często działają zgodnie z zasadą „psychologii sprawiedliwości", która czyni je wyjątkowo nieoptymalnymi w grach ekonomicznych, w których maksymalizacja indywidualnej użyteczności jest miarą sukcesu (por. Jensen, Call, Tomasello, 2007, s. 107–109). Po drugie, istnieją modele (por. Constant i in., 2019), które pozwalają wyjaśnić kooperację pomiędzy graczami ludzkimi, na podstawie innych czynników, nie czyniąc z równowagi Nasha istotnego elementu wyjaśniania. Wiąże się to bezpośrednio z odrzuceniem wielu założeń klasycznych teorii decyzji (por. podrozdział *Krytyka klasycznych modeli z perspektywy koncepcji wnioskowania aktywnego*).

to zwiększam swoją szansę na sukces moich działań, jeżeli zaś je zignoruję, to z pewnością moje działania mogą zakończyć się porażką[18].

W związku z powyższym należy wskazać, że pojęcie normatywności w odniesieniu do predykcji ma przynajmniej trzy różne znaczenia. Predykcja jest normatywna jako[19]:

- przyczyna konstytutywna;
- źródło zobowiązania;
- warunek skuteczności działań.

Przykład ten pokazuje, że istnieją określone przedmioty bądź procesy, które są normatywne ze swej natury, a nie ze względu na atrybucję dokonywaną przez obserwatora danego przedmiotu lub procesu. W tym sensie normatywność jest taką realną własnością, jaką można przypisać danemu mechanizmowi bądź przedmiotowi ze względu na funkcję, jaką on spełnia w danym organizmie, systemie poznawczym czy procesie decyzyjnym, i jakiej nie da się sprowadzić do jakichś praktyk ewaluacyjnych[20]. Innymi słowy, warunkiem przypisania normatywności jest fakt, że dany mechanizm lub przedmiot jest normatywny *per se*. Opisany przeze mnie przykład unaocznia problem, który będzie analizowany w tej pracy. Będę chciał dowieść, że aby móc wyjaśnić sukces bądź porażkę naszych działań w środowisku, należy zwrócić uwagę na określone mechanizmy biologiczne (omawiane tutaj na przykładzie mechanizmów predykcyjnych), które warunkują dobór takich, a nie innych działań. Oznacza to, że wyjaśnienie potencjalnej skuteczności naszych działań zakłada wyjaśnienie takich mechanizmów, jakie są dla tych działań normatywne. Mechanizm zaś jest normatywny wtedy, kiedy gra określoną rolę przyczynową w wyjaśnieniu takich, a nie innych działań bądź zachowań[21], a może on grać taką rolę, ponieważ odgrywa określoną rolę przyczynową w funkcjonowaniu danego mechanizmu. W analizowanym przypadku przewidywania pogody będzie zatem chodziło o rozpoznanie i opisanie mechanizmu odpowiedzialnego za wytworzenie takiego, a nie innego przewidywania, oraz wskazanie, w jaki sposób ów mechanizm pozwala wyjaśnić sukces bądź porażkę działań podejmowanych ze względu na to przewidywanie, lub takich, których to przewidywanie nie normowało[22].

---

[18] Upraszczam tutaj sytuację, bowiem może być też tak, że mam określone przewidywanie, ale z jakiegoś powodu nie uznaję go za prawdziwe. Wtedy wybór działań będzie wyglądał inaczej. Podany przeze mnie przykład jest zatem pewną idealizacją, która unaocznia określony problem.

[19] Dla uproszczenia, mówiąc o normatywności predykcji, będę miał na myśli wszystkie te trzy znaczenia.

[20] Mamy z nimi do czynienia, kiedy określamy, że np. dany organ pracuje prawidłowo. Por. analizy teleosemantycznej koncepcji funkcji w rozdziale *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego*.

[21] W znaczeniu, jakie tym pojęciom przypisuje mechanicyzm. Por. podrozdział *Przetwarzanie predykcyjne a mechanistyczny model wyjaśniania naukowego*.

[22] Niniejsza teza zakłada, że przypisanie normatywności nie jest zależne od strategii, którą Daniel Dennett (2003) wiąże z przyjmowaniem tzw. nastawienia intencjonalnego, a która polega na tym, że obserwator stara się dane zjawiska opisywać i przewidywać za pomocą przypisywanych im przekonań, pragnień czy pojęć normatywnych. W praktyce oznacza to, że korelaty tych przekonań, pragnień i pojęć wcale nie muszą istnieć. W niniejszej pracy twierdzę, że przypisanie normatywności danym przedmiotom, stanom rzeczy i procesom

Docieramy w tym miejscu do pierwszej hipotezy, której będę bronił w niniejszej pracy. Twierdzę bowiem, że:

> (H1) Należy odróżnić normatywność oceniającą[23], czyli tę, która pozwala obserwatorowi w sposób zewnętrzny (tj. z perspektywy obserwatora danego przedmiotu bądź mechanizmu) orzekać o tym, czy dany przedmiot bądź mechanizm jest skuteczny albo nieskuteczny, poprawny bądź niepoprawny; dobry albo zły itd., od normatywności wewnętrznej (perspektywa przedmiotu bądź mechanizmu), czyli tej, którą przypisuje się danemu przedmiotowi czy mechanizmowi z powodów ontycznych (czyli dlatego, że dany przedmiot czy mechanizm jest taki, jaki *jest*). Hipoteza ta nakłada na dany przedmiot bądź mechanizm warunek ontyczny.

Hipoteza ta zakłada, że normatywność (rozumiana jako własność konstytutywna dla działań i zachowań) jest realną, niezależną od obserwatora i jego praktyk ewaluacyjnych własnością danego przedmiotu, mechanizmu bądź zachowania[24].

Normatywne w tym sensie będą te własności, które są konstytutywne dla struktury i funkcjonowania danego systemu (Bickhard, 2003)[25]. Normatywność wewnętrzną można zatem nazwać „normatywnością konstytutywną" albo „normatywnością eksplanacyjną". Twierdzę, że odpowiednie ramy konceptualne, które pozwalają w sposób satysfakcjonujący opisać tak rozumianą normatywność, oferuje mechanistyczny model wyjaśnień naukowych (por. Craver, 2007; Bechtel, 2008). Zgodnie z tym modelem (1) systemy, np. poznawcze, stanowią wielopoziomowe i hierarchicznie zorganizowane układy mechanizmów; oraz (2) wyjaśnienie tych systemów polega na odkryciu i opisaniu mechanizmów z niższego poziomu organizacji, odpowiadających za realizację mechanizmów z wyższego poziomu (tj. takich, jakie konstytuują dany system). Tak rozumiany model z jednej strony wyznacza możliwość wyjaśnienia normatywności w kategoriach mechanistycznych, z drugiej zaś określa konceptualne i eksplanacyjne ramy dla rozważań przeprowadzanych w niniejszej pracy[26].

Powiązanie tak rozumianej normatywności z mechanistycznym modelem wyjaśnień w nauce ma określone konsekwencje teoretyczne. Twierdzę, i będę tego

jest możliwe dzięki temu, że przynajmniej niektóre z nich są normatywne *jako takie*. W tym sensie intencjonalność nie stanowi konstytutywnego komponentu wyjaśniania tych przedmiotów, stanów rzeczy czy procesów.

[23] Jest ona w pewnym sensie wyznaczona przez etyczne rozumienie normatywności jako tego, co określa, jakie rzeczy powinny być.

[24] W praktyce oznacza to, że normatywność jest w pewnym sensie sprowadzalna do własności funkcjonalnych lub strukturalnych określonych mechanizmów. Jest to, za czym będę argumentować, *szczególna* własność funkcjonalna.

[25] Przez system rozumiem złożony bądź prosty mechanizm, który realizuje określone funkcje. W tym sensie np. bakteria jest pewnym systemem żywym, ale również człowiek czy sztuczny system autonomiczny (np. robot albo algorytm Google'a).

[26] Na temat pojęcia mechanizmów i wyjaśniania mechanistycznego por. podrozdział *Przetwarzanie predykcyjne a mechanistyczny model wyjaśniania naukowego*.

przekonania bronił, że pojęcie normatywności ma nie tylko charakter deskryptywny, ale również eksplanacyjny. Rozumiem przez to, że jeżeli dany mechanizm[27] oznacza się jako normatywny, to ma to określone konsekwencje dla wyjaśnienia tego mechanizmu albo dla wyjaśnienia innego mechanizmu, w którym ten mechanizm gra jakieś role przyczynowe. Wszystko to składa się na drugą hipotezę, której będę tutaj bronił, a która brzmi:

> (H2) Normatywność jest pojęciem istotnym eksplanacyjnie. Za jego pomocą nie tylko opisujemy, ale także wyjaśniamy określone mechanizmy, wskazując na takie ich (normatywne) własności, jakie są dla nich konstytutywne. Hipoteza ta nakłada na dany przedmiot bądź mechanizm warunek eksplanacyjny.

Na potrzeby niniejszych badań przyjmuję założenie, że normatywne są przede wszystkim mechanizmy. Oznacza to, że normatywność można przypisać odpowiednim funkcjom, własnościom bądź całym systemom dopiero w odniesieniu do określonych mechanizmów normatywnych, które umożliwiają ich realizację, czyli są dla nich konstytutywne. Takie ustalenie może jednak budzić zrozumiałe wątpliwości. Dlaczego bowiem mówi się tutaj o mechanizmach normatywnych, a nie np. eksplanacyjnych? Przecież nie mają one charakteru reguł. Nie są również jakimiś zobowiązaniami. Na powyższą wątpliwość odpowiem, powołując się na rolę, jaką owe mechanizmy grają w wyjaśnieniach działania. Mówiąc o takich rolach w odniesieniu do normatywnych własności mechanizmów, stwierdzam, że mechanizmy te są konstytutywnymi przyczynami określonych działań danego organizmu bądź podmiotu w jego środowisku. W tym sensie można powiedzieć, że mówienie o normatywności danego mechanizmu odnosi się do faktu, że istnieją pewne zjawiska, tj. działania i zachowania, które wyjaśnia się w kategoriach mechanizmów i funkcji, które są ich przyczynami[28]. Mechanizm jest normatywny, ponieważ realizuje taką normatywną funkcję, która zobowiązuje[29] określony organizm, podmiot bądź system do wykonania takich, a nie innych działań.

---

[27] Przez mechanizm rozumiem taką strukturę, jaka pełni określoną funkcję w danym systemie dzięki swoim komponentom, ich operacjom oraz organizacji (Bechtel, 2008, s. 13).

[28] Kluczowa jest tutaj relacja pomiędzy przewidywaniem a działaniem, którą określam mianem „motywacyjnej". Przewidywanie warunkuje pojawienie się określonych działań, wykluczając tym samym inne. Oznacza to, że faworyzowane są tylko te działania, które sprzyjają i przyczyniają się do realizacji celów danego organizmu bądź podmiotu. W praktyce oznacza to, że dany organizm działa w ten sposób, który zwiększa jego szanse na przeżycie, czyli zmniejsza stopień zaskoczenia powiązany z koniecznością działania w nieznanym środowisku. Por. szczegółowe analizy tego zagadnienia w podrozdziałach *Model normatywności dla przetwarzania predykcyjnego*; *Normatywne mechanizmy predykcyjne a ograniczenia*.

[29] Pojęcie to zostanie doprecyzowane w dalszych wywodach. Trzeba jednak już teraz stwierdzić, że nie należy go traktować w kategoriach przekonań i pragnień (tj. w kategoriach psychologii potocznej) oraz myśleć o nim jako o pojęciu z teorii moralności. Odnosi się ono przede wszystkim do kategorii, które można określić jako należące do subosobowego poziomu opisu. Dlaczego zatem piszę o zobowiązaniu, a nie przyczynie? Ponieważ pojęcie zobowiązania odnosi się do pewnej dowolności w wyborze działań. Chodzi o to, że funkcja normatywna warunkuje pojawienie się określonych działań, wykluczając w ten sposób inne.

W tym sensie owa normatywna funkcja, realizowana przez określony mechanizm normatywny, stanowi konstytutywną przyczynę takiego, a nie innego działania.

Broniony tutaj pogląd w pewnym stopniu koresponduje z ujęciami normatywności zaproponowanymi przez takich autorów jak Ruth Millikan czy Mark Bickhard. Oni również uważają, że istnieją pewne funkcje czy własności normatywne, za pomocą których można wyjaśniać funkcjonowanie takich, a nie innych systemów. Jednak stawką, o którą toczy się gra, jest wykazanie, która z hipotez o normatywności jest zasadna. Czy hipoteza ontyczna głosząca, że istnieją określone mechanizmy, funkcje bądź własności w świecie, które można nazwać „normatywnymi"? Czy może raczej hipoteza epistemiczna, która głosi, że posiadamy dobre racje (pragmatyczne, eksplanacyjne itd.), aby niektórym mechanizmom, funkcjom czy własnościom przypisać normatywność? W niniejszej pracy będę bronił hipotezy ontycznej i jednocześnie twierdził, że uzasadnia ona hipotezę eksplanacyjną. Tym sposobem docieramy do trzeciej hipotezy, która w pewnym sensie ugruntowuje dwie pozostałe – (H1) i (H2):

> (H3) Jeżeli istnieją mechanizmy normatywne, to grają one określone role przyczynowe w wyjaśnieniach działań i zachowań[30]. Hipoteza ta nakłada na dany przedmiot bądź mechanizm warunek normatywności (eksplanacyjnej).

Trzy wymienione przeze mnie hipotezy, jeżeli zostaną dobrze uzasadnione, mogą prowadzić do sformułowania ogólnej teorii mechanizmów biologicznych[31]. Celem tej pracy będzie wykazanie, że taka teoria jest możliwa. Jednak samo jej zbudowanie znacznie wykracza poza ramy prezentowanych tutaj analiz. Moje zadanie będzie dużo skromniejsze. Na przykładzie ramy[32] przetwarzania predykcyjnego (dalej w tekście będę ją oznaczał jako PP[33], od *predictive processing*) (por. Clark, 2013b, 2016; Hohwy, 2013, 2014, 2016; Wiese, Metzinger, 2017), stanowiącej jedną z najżywiej obecnie dyskutowanych propozycji rozumienia poznania, będę chciał wykazać, że (1) broniona przeze mnie teoria mechanizmów normatywnych może zostać zastosowana w kognitywistyce i filozofii biologii; oraz że (2) rozpoznanie i omówienie opisywanych przez tę ramę badawczą mechanizmów, które można

[30] W rozdziale *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego* stwierdzę, że satysfakcjonująca teoria normatywności dla funkcji i mechanizmów musi spełniać dwa wymogi: normatywny i eksplanacyjny.

[31] Przez mechanizm biologiczny rozumiem system bądź układ przyczynowo oddziałujących części (komponentów) i procesów (operacji), które wywołują jeden lub więcej efektów.

[32] Korzystam z pojęcia ramy (*framework*), ponieważ (1) PP nie jest (jeszcze) określoną propozycją teoretyczną, a zbiorem kilku koncepcji oraz (2) wskazuje ono raczej potencjalne kierunki badań, aniżeli określa takie, a nie inne strategie eksplanacyjne (Fink, Zednik, 2017). Na kartach tej pracy będę bronił twierdzenia, zgodnie z którym rama ta pozwala w sposób nietrywialny modelować czynności poznawcze, niemniej w punkcie wyjścia wolę przypisać jej mniejsze możliwości eksplanacyjne.

[33] W tekście niniejszej pracy ten skrót będzie stosowany w różnych kontekstach. Będę zatem wspominał o „ramie PP", „modelu PP" czy „koncepcji PP". Zwroty te należy czytać jako „rama przetwarzania predykcyjnego", „model przetwarzania predykcyjnego" itd.

uznać za normatywne, ma nietrywialne znaczenie eksplanacyjne dla badań przeprowadzanych w tej ramie.

Przetwarzanie predykcyjne (PP) jest inspirowaną biologicznie i neurologicznie koncepcją ludzkiego poznania, oferującą jego wyjaśnienie w terminach obliczeń bayesowskich. Zgodnie z tym ujęciem żywy organizm ucieleśnia[34] albo wręcz jest[35] wielopoziomowym, hierarchicznym modelem generatywnym swojego środowiska[36], które przechowuje bądź przetwarza informacje w sposób hierarchiczny[37]: wysyła on odgórnie generowane przewidywania (tj. predykcje), których celem jest zminimalizowanie oddolnego sygnału o tzw. błędzie predykcyjnym. Błąd ten odnosi się do dysproporcji pomiędzy oczekiwaniami opartymi na wewnętrznych parametrach modelu a zmiennymi informacjami docierającymi do modelu za pośrednictwem zmysłów. W myśl tego poglądu organizm może skutecznie działać w środowisku, jeżeli potrafi sprawnie zminimalizować potencjalne błędy predykcyjne. Opisany tutaj schemat obliczeniowy ma umożliwić wyjaśnienie percepcji, działania, poznania, uwagi itd. (por. Clark, 2013b; Hohwy, 2013, 2015a; Seth, 2015).

Wydaje się, że PP jest dobrym przykładem koncepcji, w której mechanizmy normatywne można wskazać i opisać. Powodów tego stanu rzeczy jest kilka. Po pierwsze jest to rama badawcza, która obiecuje w sposób naturalistyczny i biologicznie wiarygodny wyjaśnić złożone interakcje na linii percepcja – działanie. Po drugie, jeżeli wierzyć prominentnym przedstawicielom tego stanowiska, to PP może zaoferować unifikację teorii w kognitywistyce (por. Clark, 2013b; Hohwy, 2015a; Seth, 2015)[38]. Po trzecie PP oferuje wyjaśnianie percepcji i działania, wykorzystując modelowanie bayesowskie, co w praktyce oznacza, że system poznawczy ma (w jakimś stopniu) implementować bayesowską sieć przekonań (por. Harkness, Keshava, 2017; Hohwy, 2013). Po czwarte wielu badaczy (por. Friston, Kiebel, 2009; Hohwy, 2014, 2015; Seth, 2015) uważa, że PP wywodzi się z bardziej podstawowej, biologicznej teorii życia opartej na tzw. zasadzie energii swobodnej

[34] Pojęcie modelu generatywnego jest sporne. Niektórzy sugerują, że model generatywny należy rozpatrywać w odniesieniu do całego organizmu (por. Bruineberg, Rietveld, 2014; Bruineberg, Kiverstein, Rietveld, 2016; Friston, 2013a). Inni twierdzą, że nie jest do końca jasne, czy dotyczy to każdego żywego organizmu, czy tylko ludzi oraz małp naczelnych (por. Hohwy, 2020a, s. 11–12). W niniejszej pracy sympatyzuję jednak ze stanowiskiem, które wiąże lokalizację modelu generatywnego z mózgiem i przypisuje jego posiadanie tylko ludziom i naczelnym (por. podrozdziały *Konserwatywne przetwarzanie predykcyjne – Pośredniość percepcji w przetwarzaniu predykcyjnym*). Nie będę jednak tutaj rozstrzygał tej kwestii szczegółowo, ponieważ znacznie wykracza ona poza rozważania, które są celem tej pracy.

[35] Pojęcie modelu generatywnego jest bardzo sporne. Niektórzy sugerują, że model ten należy rozpatrywać w odniesieniu do całego organizmu (por. Bruineberg, Rietveld, 2014; Bruineberg, Kiverstein, Rietveld, 2016; Friston, 2013a).

[36] Ogólnie mówiąc, mózg w konserwatywnej wersji PP (por. podrozdział *Konserwatywne przetwarzanie predykcyjne*) rozumiany jest jako przyczynowo probabilistyczny model świata. Pojęcie modelu świata nie jest nowe w filozofii. Zaproponował je w swoich pracach Andrzej Półtawski (1973, 1986), inspirując się książką francuskiego psychiatry Henri Ey'a pt. *La conscience* (1956). Por. także prace uczniów Półtawskiego: Maciejczak (2008); Piłat (1999).

[37] Więcej na temat pojęcia modelu generatywnego por. podrozdział *Model generatywny*.

[38] Andy Clark (2013b) określa ją jako *grand unified theory* (GUT), czyniąc wyraźne aluzje do fizycznej teorii wielkiej unifikacji.

(*free energy principle* – w dalszej części będę ją oznaczał jako FEP). Zgodnie z tą zasadą organizmy żywe są systemami, które utrzymują swoje istnienie na drodze minimalizowania energii swobodnej swoich stanów wewnętrznych (por. Friston, Stephan, 2007; Friston 2009, 2010, 2012, 2017). W myśl poglądów jej zwolenników FEP prowadzi nie tylko do satysfakcjonującej teorii wyjaśniającej życie i poznanie, ale także stanowi zasadę normatywną (por. Adams, Brown, Friston, 2014; Allen, Friston, 2018; Friston, 2013c; Friston, Fortier, Friedman, 2018), ponieważ orzeka m.in., że pewne rzeczy są z nią zgodne, a inne nie; a także dostarcza normatywny opis, zgodnie z którym mózg bądź żywy organizm, aby mógł zachować swoją własną integralność, *musi* angażować się w minimalizowanie energii swobodnej, *resp*. błędów predykcyjnych (Allen, Friston, 2018, s. 2472). W tym sensie FEP jest zasadą, która pokazuje, dlaczego minimalizacja energii swobodnej, *resp*. błędów predykcyjnych, ma być warunkiem możliwości istnienia danego systemu (por. Colombo, Wright, 2018, s. 4).

Twierdzę zatem, że PP stanowi taką ramę badawczą, w jakiej można w sposób nietrywialny wskazać i opisać mechanizmy normatywne, spełniające warunki określone przez trzy bronione w tej pracy hipotezy, tj. warunek ontyczny, warunek eksplanacyjny oraz warunek normatywności (eksplanacyjnej). W odniesieniu do PP będę zatem bronił następującej hipotezy:

(H4) PP oferuje nietrywialne wyjaśnienie określonych zjawisk, wykorzystując modelowanie Bayesa.

Aby móc przyjąć prawdziwość tej hipotezy, należy wpierw uzasadnić prawdziwość jednej z dwóch hipotez pomocniczych. Pierwszą z nich określam mianem „hipotezy epistemicznej". Głosi ona, że:

(HE) Istnieją takie zjawiska, jakie wyjaśnia się za pomocą bayesowskiego PP z powodów pragmatycznych, instrumentalnych, eksplanacyjnych itd.

Hipotezie tej przeciwstawiam hipotezę ontyczną, która głosi, że:

(HO) Istnieją takie zjawiska, jakie wyjaśnia się za pomocą bayesowskiego PP, ponieważ są one bayesowskie z natury, a jeżeli są one bayesowskie z natury[39], to znaczy, że powinno się je wyjaśniać przy użyciu modelowania probabilistycznego.

[39] Zwrot „z natury" zapożyczam w pewnym stopniu z dyskusji nad systemami dynamicznymi (por. Chemero, 2014; Van Gelder, 1995). Odróżnia się tutaj dwa rodzaje tez: ontyczne i epistemiczne. Te pierwsze zakładają, że naturalne systemy korzystają/bądź nie korzystają np. z reprezentacji. Co oznacza, że wewnątrz nich istnieje bądź nie istnieje coś, co spełnia warunki określane przez taką, a nie inną teorię reprezentacji. Innymi słowy, można powiedzieć, że zgodnie z tezą ontyczną dany system jest bądź nie jest reprezentacyjny z natury, tzn. jest taki, bo taki jest. Te drugie zaś zakładają, że nasze modele lub wyjaśnienia mają się odwoływać lub nie mają się odwoływać (zakładać ich istnienie bądź ich nieistnienie) np. do reprezentacji. Tezy te

Celem niniejszej pracy jest uzasadnienie hipotezy ontycznej oraz wykazanie, że (przynajmniej niektóre) mechanizmy opisywane przez PP są normatywne. W związku z tym bronię jeszcze jednej hipotezy, która jest dedykowana specjalnie PP:

(H5) Bayesowskie PP ma charakter normatywny, ponieważ mechanizmy predykcyjne są normatywne (w sensie przedstawionym powyżej), a nie dlatego, że formułuje się tutaj pewne (normatywne) zasady działania lub myślenia.

Hipotezy czwarta i piąta stanowią egzemplifikację pierwszych trzech hipotez. Wszystkie te hipotezy prowadzą do wykazania, że bayesowskie PP jest (1) nietrywialną ramą wyjaśniającą (przynajmniej niektóre) zjawiska poznawcze; oraz (2) ramą normatywną, ponieważ wyjaśnia zjawiska poprzez rekonstrukcję i opis mechanizmów normatywnych odpowiedzialnych za ich realizację, a nie dlatego, że formułuje jedną lub więcej zasad normatywnych[40]. Inaczej mówiąc: bayesowskie PP jest normatywne, ponieważ (przynajmniej niektóre) zjawiska przezeń wyjaśniane są normatywne z natury. Takie ujęcie pozwala na obronę PP przed zarzutami formułowanymi przez zwolenników teorii empirycznych wobec ujęć preskryptywnych. Normatywność PP wywodzi się z normatywności mechanizmów, które spełniają określone funkcje przyczynowe. Nie ma tutaj mowy o żadnych regułach bądź normach, które stanowić miałyby swoistą miarę poprawności, skuteczności bądź optymalności określonych działań i zachowań.

określają, że dany system jest bądź nie jest reprezentacyjny ze względu na możliwą zewnętrzną (względem tego systemu) atrybucję danej cechy lub własności (por. także podrozdział *Ujęcie nonkognitywistyczne* a ujęcie *kognitywistyczne*). Mówiąc zatem, że dane zjawisko jest bayesowskie z natury, mam na myśli wyżej wyeksplikowane rozumienie wyrażenia „z natury".

[40] Por. rozdział *Mechanizmy normatywne a działania w przetwarzaniu predykcyjnym*, w którym uzasadniam niniejszą tezę.

Przedmowa

Celem niniejszej pracy jest uzasadnienie wszystkich pięciu (wcześniej postawionych przeze mnie) hipotez oraz hipotezy ontycznej, która wspiera hipotezy czwartą i piątą[1]. Innymi słowy, będę bronił przekonania, zgodnie z którym istnieją biologiczne mechanizmy normatywne, które spełniają nietrywialne role przyczynowe w formułowanych przez nas wyjaśnieniach działań i zachowań określonych systemów. Przykładem takich mechanizmów są opisywane i wyjaśniane przez PP mechanizmy predykcyjne, które (1) przewodzą działaniom oraz (2) kształtują przyczynowe przejścia pomiędzy stanami, które posiadają określoną treść i warunki spełniania (np. stany mentalne). Przyświeca mi zatem określony cel teoretyczny, który wiąże się z koniecznością wskazania tych warunków, które *powinna* spełniać nietrywialna teoria mechanizmów normatywnych oraz określone modele proponowane przez zwolenników PP.

W niniejszej pracy zostaną wykorzystane klasyczne metody filozoficzne, takie jak analiza pojęć czy krytyczny namysł. Będę analizował wybrane badania z zakresu kognitywistyki, psychologii poznawczej, neurologii, teorii informacji i biologii w aspekcie wykorzystanej metodologii, argumentacji i języka, dobierając je ze względu na ich teoretyczną ważność dla omawianych w tej pracy zagadnień. W tym sensie prezentowane tutaj badania mają charakter interdyscyplinarny. Odpowiednią ramę badawczą wyznaczy mi mechanistyczny model wyjaśniania w nauce, który określa konieczne i wystarczające warunki wyjaśnienia danego zjawiska. Obrane przeze mnie metody badawcze są zatem zrelatywizowane do problemów, które zamierzam rozwiązać.

W tym miejscu chciałbym przedstawić układ treści niniejszej rozprawy. Wprowadzający rozdział *Pojęcie przetwarzania predykcyjnego* poświęcony będzie omówieniu tego, czym jest PP oraz jakie są jego główne założenia i tezy. Zostaną wprowadzone kluczowe dla tej ramy badawczej pojęcia oraz dystynkcje. Wielu autorów twierdzi, że PP jest współczesną wersją kantyzmu i narażone jest na podobne zarzuty, jakie można stawiać koncepcji Immanuela Kanta. Przedyskutuję tę tezę i wykażę, że tylko w bardzo ogólnym sensie rama PP ma charakter neokantowski. Nie mamy tutaj bowiem do czynienia ani z dedukcją transcendentalną, ani z za-

[1] Należy podkreślić, że pojęcia hipotezy i uzasadniania są pojęciami pragmatycznymi. W przypadku uzasadniania dedukcyjnego dowód hipotezy, tj. zdania, o którym ktoś przypuszcza, że jest prawdziwe – zmienia tę hipotezę po prostu w tezę lub pozwala do teorii dołączyć określoną implikację, której poprzednikiem jest owa hipoteza. Jednak wszelkie niededukcyjne uzasadnienia zwiększają po prostu prawdopodobieństwo tego, że postawiona hipoteza jest prawdziwa. Można zatem powiedzieć, że ona jest wciąż hipotezą tylko prawdopodobniejszą niż poprzednio. W uzasadnieniu niededukcyjnym to prawdopodobieństwo nigdy nie jest równe 1, co jest istotą uzasadnień niededukcyjnych (por. Mortimer, 1982).

stosowaniem argumentów transcendentalnych. Będę twierdził, że PP oparte jest na metodzie inżynierii odwrotnej i wnioskowaniach abdukcyjnych. W drugiej części tego rozdziału odpowiem na zarzut sformułowany przez Dana Zahaviego, który wprost zarzuca tej ramie badawczej konsekwencje antyrealistyczne. Wykażę, że stanowisko internalizmu, obecne w tzw. konserwatywnym PP, nie implikuje antyrealizmu, a z powodu eksplanacyjnej roli, jaką grają w nim skierowane na realne wzorce S-reprezentacje (reprezentacje strukturalne), zasadne jest twierdzenie, że PP ma charakter realistyczny. Tym sposobem wykażę, że PP to nietrywialna rama badawcza, mająca swój przedmiot, określone metody oraz właściwy sobie status epistemiczny. Na koniec omówię stanowiska zaliczane do tzw. radykalnego PP.

W rozdziale *Przetwarzanie predykcyjne jako bayesowski model eksplanacyjny* będę uzasadniał tezę, zgodnie z którą PP oferuje modelowanie typu bayesowskiego. Wielu badaczy twierdzi bowiem, że mózg jest zaimplementowaną statystyczną siecią probabilistyczną, która w *jakimś* sensie realizuje regułę Bayesa, co w praktyce ma oznaczać, że wszelkie procesy poznawcze mają być zastosowaniem reguły Bayesa i dają się opisać w kategoriach rozkładów prawdopodobieństwa. Takie rozwiązanie wzbudza u wielu badaczy sprzeciw i jest przedmiotem szerokiej krytyki. Celem tego rozdziału jest uzasadnienie, że bayesowskie PP jest nietrywialną ramą badawczą. W tym celu będę argumentował za tym, że wyjaśnia ono określone zjawiska nie tylko na opisanym przez Davida Marra poziomie obliczeniowym, ale także na poziomach algorytmów i implementacji. W dalszej części rozdziału będę argumentował za tym, że PP jest modelowaniem normatywnym. Zwolennicy wykorzystania modeli bayesowskich w psychologii czy teorii decyzji twierdzą, że są one normatywne, ponieważ pozwalają na sformułowanie formalnych reguł działania, które pokazują, co należy robić, aby dane działanie było optymalne. Krytycy tego podejścia podkreślają, że takie myślenie o normatywności modelowania bayesowskiego jest nieuzasadnione oraz że w nauce należy przejść z pozycji preskryptywnych na deskryptywne. W polemice z Shirą Elqayam i Jonathanem Evansem (2011) wykażę, że proponowany przez nich podział na preskryptywizm i deskryptywizm bayesowski jest pozorny, ponieważ ma się tutaj do czynienia z dwiema formami preskryptywizmu, tj. słabą i mocną. Wykażę, że wersja słaba ma charakter epistemiczny i może prowadzić do antyrealizmu, zaś mocna jest ontyczna i pozwala uzasadnić realizm w odniesieniu do modeli bayesowskich. Będę twierdził, że słaba wersja preskryptywizmu jest zasadna w odniesieniu do PP. Pozwala ona przyjąć antyrealizm w stosunku do PP. W praktyce oznacza to, że można wyjaśniać zjawiska przy użyciu twierdzenia Bayesa. Nie implikuje to jednak tezy, że są one bayesowskie z natury. Pełne uzasadnienie realizmu w odniesieniu do bayesowskiego PP zakłada jednak przyjęcie mocnego preskryptywizmu. Stanowisko to zakłada, że zjawiska wyjaśnia się za pomocą twierdzenia Bayesa, ponieważ są one bayesowskie. Jeżeli są one bayesowskie z natury, to znaczy, że *powinno* się je wyjaśniać przy użyciu

modelowania Bayesa. Teza ta zostanie uzasadniona w rozdziałach *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego* i *Mechanizmy normatywne a działania w przetwarzaniu predykcyjnym*. W rozdziale tym zatem wstępnie uzasadniam hipotezy pierwszą (H1), drugą (H2) oraz epistemiczną (HE), wskazując jednocześnie na konieczność uzasadnienia hipotezy ontycznej (HO).

W rozdziale *Zasada energii swobodnej w przetwarzaniu predykcyjnym* omówię sformułowaną przez Karla Fristona FEP. Zgodnie z tą zasadą wszystkie systemy biologiczne (zdefiniowane w terminach koców Markowa) minimalizują energię swobodną swoich stanów wewnętrznych w celu utrzymania homeostazy. Niektórzy badacze uważają, że PP jest specjalnym przypadkiem zastosowania tej zasady odniesionym do poznania, a mechanizmy predykcyjne są homeostatycznymi mechanizmami minimalizującymi energię swobodną. Omówienie FEP jest istotne ze względu na to, że część autorów uważa ją za istotną eksplanacyjnie i normatywną. Jeżeli tak jest, to FEP okazuje się kluczowa dla wyjaśnienia normatywnych mechanizmów predykcyjnych i w ogóle wszelkich normatywnych mechanizmów biologicznych. Aby określić możliwości eksplanacyjne tej zasady, odwołam się do dyskusji jej zwolenników na temat zagadnienia określanego przez nich jako problem ciągłości pomiędzy życiem a umysłem (*continuity and discontinuity between life and mind*). Krytyczna analiza tej dyskusji oraz dodatkowe, sformułowane przeze mnie argumenty pozwolą mi na rewizję pretensji eksplanacyjnych FEP. Odrzucę również przekonanie, zgodnie z którym zasada ta jest konieczna dla wyjaśnienia natury mechanizmów predykcyjnych. Będę argumentował za tym, że zasada sformułowana i broniona przez Fristona stanowi jedynie istotną heurystykę badawczą dla analiz z zakresu PP.

W rozdziale *Funkcje i mechanizmy normatywne w przetwarzaniu predykcyjnym* moje analizy rozpocznę od sformułowania odpowiedzi na pytanie dotyczące normatywnej natury mechanizmów homeostatycznych. Wykażę, że mechanizmy predykcyjne nie są homeostatyczne. Będę bronił poglądu, że pełne wyjaśnienie mechanizmów normatywnych zakłada wyjaśnienie funkcji normatywnych. Przedyskutuję najważniejsze propozycje rozumienia normatywności funkcji, zarówno w perspektywie systemowej, jak i teleosemantycznej. Będę twierdził, że nietrywialna koncepcja funkcji musi spełniać dwa wymogi, które określam jako eksplanacyjny i normatywny. Wykażę, że żadna z przywołanych przeze mnie teorii nie spełnia satysfakcjonująco obu tych wymogów. Zaproponuję model normatywności oparty na koncepcji Bickharda, ale uzupełniony o perspektywę mechanistyczną. Będę twierdził, że funkcja jest normatywna wtedy, gdy jednocześnie: (1) pozwala wyjaśnić dysfunkcję danego mechanizmu; (2) przyczynia się do utrzymania stabilności organizmu w ten sposób, że kształtuje i ogranicza możliwe relacje, procesy i zachowania danego systemu; oraz gdy (3) (w odniesieniu do funkcji reprezentowania i predykcyjnej) pozwala wyjaśnić możliwość przypisywania wartości

logicznych określonym reprezentacjom/predykcjom. W takim ujęciu mechanizm jest normatywny wtedy, gdy realizuje określone funkcje normatywne i gdy jest konstytutywny dla określonej aktywności, pomimo faktu, że z jakiegoś powodu nie może jej obecnie lub w dłuższej perspektywie realizować.

Takie rozumienie normatywności mechanizmów zakłada przyjęcie hipotezy epistemicznej (HE). Będę twierdził, że hipoteza ta nie jest satysfakcjonująca poznawczo, w związku z czym należy uzasadnić hipotezę ontyczną (HO), co bezpośrednio wiąże się z przyjęciem stanowiska preskryptywizmu ontycznego. Z tego też powodu, w odniesieniu do mechanistycznej teorii wyjaśnień naukowych, sformułuję ontyczną interpretację pojęcia mechanizmu normatywnego. Zgodnie z tym ujęciem mechanizm albo funkcja są normatywne wtedy, gdy spełniają one takie, a nie inne funkcje przyczynowe w wyjaśnieniach określonych działań i zachowań. W odniesieniu do normatywnych własności mechanizmów i funkcji predykcyjnych oznacza to, że są one przyczynami określonych działań danego organizmu w środowisku. Tym sposobem potwierdzę hipotezę drugą (H2), trzecią (H3), a także wstępnie hipotezę piątą (H5). Uzasadnię również konieczność przyjęcia hipotezy ontycznej (HO) i odrzucenia hipotezy epistemicznej (HE).

Rozdział piąty zatytułowany *Mechanizmy normatywne a aktywne wnioskowanie w przetwarzaniu predykcyjnym* poświęcony będzie problemowi ciemnego pokoju i powiązanemu z nim kompromisowi eksploracji i eksploatacji. Ciemny pokój oznacza stan, w którym mógłby znaleźć się podmiot, gdyby zminimalizował sumę wszystkich potencjalnych błędów predykcyjnych. Pokażę, że zgodnie z bazowym założeniem PP o konieczności ciągłej i długoterminowej minimalizacji błędów predykcyjnych, taki stan powinien być dla podmiotu pożądany. Czy tak faktycznie jest? Wielu autorów sądzi, że nie. Będę argumentował za tym, że sprawdzianem wartości PP oraz FEP jest możliwość nietrywialnego rozwiązania tego problemu, który można sprowadzić do wyboru pomiędzy aktywną i zwiększającą niepewność eksploracją a bezpieczną i łatwo przewidywalną eksploatacją. Wykażę, że obecne w literaturze rozwiązanie zaproponowane przez zwolenników PP i FEP nie pozwala na w pełni satysfakcjonujące wyjaśnienie tego kompromisu.

Następnie będę bronił stanowiska, zgodnie z którym pełne wyjaśnienie mechanizmów normatywnych, a w dalszej kolejności rozwiązanie dylematu eksploracji i eksploatacji, zakłada odwołanie się do istnienia ograniczeń obecnych w środowisku. Ograniczenia są m.in. tym elementem otoczenia, który czyni dany mechanizm nie tylko przyczynowym, ale także normatywnym. Są one zatem kluczowe dla wyjaśnienia mechanizmów predykcyjnych. Nie pełnią one bowiem tylko funkcji kontekstu, w którym mechanizm jest realizowany, ale przede wszystkim są jego konstytutywnym komponentem. Będę twierdził, że pełne wyjaśnienie roli ograniczeń w normatywnych mechanizmach predykcyjnych zakłada integrację poszczególnych modeli określonych zjawisk poznawczych, ponieważ dopiero me-

chanistyczna integracja PP z innymi modelami pozwala na nietrywialne i mocne eksplanacyjnie wyjaśnienie natury normatywnych mechanizmów predykcyjnych. Monizm eksplanacyjny obecny w wielu ujęciach PP przesądza o niemożliwości rozwiązania problemu ciemnego pokoju.

W dalszej części tego rozdziału będę argumentował za tym, że bayesowskie PP jest normatywne nie dlatego, że umożliwia sformułowanie takich, a nie innych reguł postępowania, ale dlatego, że normatywne są same mechanizmy predykcyjne. Są one normatywne, ponieważ warunkują wybór takich, a nie innych działań przez podmioty. Tym sposobem uzasadnię hipotezę piątą (H5). Hipoteza czwarta (H4) zostanie uzasadniona poprzez wykazanie, że mechanizmy normatywne pozwalają na wyjaśnienie, kluczowego dla rozwiązania problemu ciemnego pokoju, zjawiska motywacji podmiotów. Tym sposobem zostaną uzasadnione wszystkie hipotezy, a cel książki zostanie zrealizowany.

W ostatniej części rozdziału sformułuję hipotezę normatywności rozproszonej, która zakłada, że o normatywnym charakterze określonych mechanizmów, funkcji czy przedmiotów decydują relacje, w które te mechanizmy, funkcje czy przedmioty wchodzą. Oznacza to, że normatywna (w sensie pierwotnym) jest struktura relacyjna, która konstytuuje normatywność określonych przedmiotów wchodzących w jej skład. Zasugeruję, że hipoteza ta otwiera przed badaczami wiele obszarów badań i umożliwia przemyślenie wielu problemów na nowo.

W *Zakończeniu* podsumuję wyniki moich badań oraz wskażę dalsze perspektywy badawcze.

* * *

Pomysł na tę książkę rodził się długo. Przekonanie o normatywnym charakterze mechanizmów predykcyjnych towarzyszyło mi właściwie od samego początku mojego zainteresowania PP. Jednak satysfakcjonujące sformułowanie tego przekonania wymagało ode mnie wiele czasu i pracy. Sam z pewnością nie doszedłbym do myśli i tez sformułowanych w tej książce i w tej formie, gdyby nie pomoc kilku osób, którym za inspirujące rozmowy, liczne rady, lektury kolejnych fragmentów tej pracy chcę tu podziękować. Słowa wdzięczności należą się prof. Marcinowi Miłkowskiemu, dr. Pawłowi Gładziejewskiemu, dr. Witoldowi Wachowskiemu, Tomaszowi Korbakowi, dr. Maxwellowi Ramsteadowi, dr. Przemysławowi Nowakowskiemu, dr. Andrzejowi Waleszczyńskiemu oraz Michałowi Obidzińskiemu. Ich pomoc była znacząca, niemniej cała odpowiedzialność za wszelkie niedociągnięcia, błędy tej książki oraz jej słabości spoczywa na mnie.

Szczególne podziękowania należą się Dziekanowi Wydziału Filozofii Chrześcijańskiej i mojemu bezpośredniemu przełożonemu ks. prof. Janowi Krokosowi, który wspierał mnie na każdym etapie pisania tej książki oraz zawsze służył mi radą, szczególnie w chwilach największych wątpliwości.

Pragnę również podziękować władzom Uniwersytetu im. Kardynała Stefana Wyszyńskiego w Warszawie za możliwość przejścia na etat naukowy w roku akademickim 2018/2019, co bezpośrednio przyczyniło się do szybszego napisania tej książki. Moja macierzysta Uczelnia służyła mi również pomocą finansową, dzięki której poszczególne fragmenty niniejszej pracy mogłem dyskutować na konferencjach oraz seminariach. Niektóre badania, które umożliwiły mi napisanie wielu fragmentów tej pracy, zostały sfinansowane ze środków Narodowego Centrum Nauki przyznanych na podstawie decyzji numer DEC-2017/01/X/HS1/00165.

Dodatkowe podziękowania należą się studentom Instytutu Filozofii Uniwersytetu Kardynała Stefana Wyszyńskiego, którzy uczęszczali na moje wykłady poświęcone PP. Ich pytania oraz wątpliwości niejednokrotnie zmuszały mnie do przemyślenia najbardziej podstawowych kwestii. Dziękuję również uczestnikom organizowanego przez prof. Marcina Miłkowskiego w Instytucie Filozofii i Socjologii Polskiej Akademii Nauk seminarium *Filozofia kognitywistyki* za inspirujące dyskusje nad fragmentami niniejszej pracy.

Ponadto dziękuję recenzentom wydawniczym tej książki, tj. prof. dr hab. Adrianie Schetz i prof. dr. hab. Pawłowi Kawalcowi, którego uwagi dotyczące bayesizmu i wielu tez głoszonych w niniejszej pracy znacząco przyczyniły się do zachowania jasności moich wywodów. Wierzę, że udało mi się zasadnie odpowiedzieć na postawione przez niego pytania.

Na koniec chciałbym słowa podziękowania skierować do tych, którzy ponieśli największy koszt pisania tej pracy, czyli do mojej rodziny. Jestem wdzięczny mojej żonie Agnieszce za nieustanne wspieranie mnie, cierpliwość i stworzenie warunków, w których mogłem spokojnie pracować. Dziękuję moim dzieciom: Kajetanowi, Kornelii, Zuzannie, Maksowi i Jasiowi, że pozwolili mi na to, abym czas, który mogłem spędzić z nimi, wielokrotnie wykorzystał na pisanie tej pracy. Żonie i dzieciom dedykuję niniejszą książkę. Dodatkowe podziękowania należą się moim rodzicom Katarzynie i Andrzejowi oraz bratu Maciejowi za wiarę w moje siły i wsparcie.

Tekst niniejszej pracy nie był dotąd publikowany. Wyjątkiem są pewne fragmenty *Wprowadzenia*, które w zmienionej postaci ukazały się w *Studiach Philosophiae Christianae* (Piekarski, 2018), oraz rozdziału *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego*, który w rozbudowanej i zmodyfikowanej postaci ukazał się w *Avant. Pismo Awangardy Filozoficzno-Naukowej* (Piekarski, 2020). Dziękuję redakcjom obu czasopism za zgodę na wykorzystanie wskazanych prac.

Rozdział 1
Pojęcie przetwarzania predykcyjnego

Wstęp

W niniejszym rozdziale wprowadzam podstawowe pojęcia związane z PP oraz przedstawiam ogólną charakterystykę tej ramy badawczej. Wskazuję na filozoficzne i pozafilozoficzne źródła tej ramy, odnosząc się do przekonania, że PP jest koncepcją neokantowską. Stwierdzam, że w sensie ogólnym można je uważać za kantowskie, niemniej mamy tutaj do czynienia z inną metodologią, która nie jest oparta na dedukcji transcendentalnej, ale na metodzie inżynierii odwrotnej i wnioskowaniu abdukcyjnym. W dalszej części rekonstruuję zarzut antyrealizmu sformułowany przez Zahaviego. Autor ten twierdzi, że przyjmując założenia PP, jesteśmy zmuszeni porzucić postawę realistyczną. Odpowiadając na ten zarzut, wykażę, że internalizm obecny w PP (który zdaniem Zahaviego implikuje neurocentryzm i antyrealizm) ma charakter epistemiczny, tzn. granica systemu poznawczego jest ustalana w zależności od tego, gdzie umieści się koc Markova. O realistycznym charakterze PP przesądza fakt, że przynajmniej niektóre wzorce statystyczne w sygnale zmysłowym mają określoną siłę dowodową, ponieważ odpowiadają im realne struktury w świecie. W tym sensie są to wzorce realne. Za ich pośrednictwem model ma dostęp do określonych stanów świata. Następnie wykażę, że model generatywny reprezentuje owe realne wzorce i wykorzystuje S-reprezentacje. Oznacza to, że nawet jeżeli, jak twierdzi część badaczy, S-reprezentacje są neuroreprezentacjami, to ich istnienie nie prowadzi do rozstrzygnięć antyrealistycznych, co sugeruje Zahavi, tylko realistycznych.

W dalszej części tego rozdziału, wykorzystując ujęcie Zoe Drayson, będę twierdził, że w PP można mówić o bezpośredniości percepcji w trzech niesprowadzalnych sensach, tj. psychologicznym, metafizycznym i epistemologicznym. Będę argumentował za tym, że tylko w sensie psychologicznym można twierdzić, że percepcja w PP jest pośrednia. Nie oznacza to jednak, że pośredniość psychologiczna wyklucza automatycznie bezpośredniość metafizyczną bądź epistemologiczną percepcji. Na koniec scharakteryzuję stanowiska składające się na tzw. radykalne PP oraz wskażę ich genezę i specyfikację.

1.1. Ogólna charakterystyka i podstawowe pojęcia

Moje rozważania rozpocznę od ogólnego wyjaśnienia tego, czym jest PP. Takie omówienie pozwoli czytelnikowi na w miarę swobodne poruszanie się w obrębie omawianych tu problemów. Wraz z kolejnymi rozważaniami prezentowane tutaj rozumienie PP będzie doprecyzowywane i zyska na jasności. Na wstępie należy stwierdzić, że nie ma jednej ogólnie przyjętej koncepcji PP. Nie ma również jednego terminu, którym tę koncepcję się określa. W literaturze można spotkać takie określenia jak proponowana przez Jakoba Hohwy'ego „minimalizacja błędu predykcyjnego" (*predictive error minimization*) czy „kodowanie predykcyjne" (*predictive coding*), którym najchętniej posługują się neuronaukowcy (por. m.in. Rao, Ballard, 1999; Spratling, 2017). Z pierwszego z tych określeń rezygnuję, ponieważ ściśle wiąże się ono z wykładnią PP proponowaną przez Hohwy'ego. Nie pokrywa się z ujęciem tu rozwijanym, pomimo że wiele mu ono zawdzięcza. Z kodowania predykcyjnego nie chcę zaś korzystać, ponieważ odnosi się ono głównie do takich strategii kodowania, w jakich przewidywane części kodu sygnału wejściowego są odejmowane od rzeczywistego sygnału. W ten sposób tylko różnica pomiędzy tymi sygnałami jest przekazywana jako dana wyjściowa do następnego etapu przetwarzania informacji. W takich ujęciach mniejszą uwagę przywiązuje się do idei hierarchicznego i wielopoziomowego modelu generującego predykcje (Clark, 2016, s. 25–26; Kiefer, Hohwy, 2017, s. 9)[1].

Autorzy rozwijający różne ujęcia i stanowiska dotyczące PP podzielają, w większym lub mniejszym stopniu, następujące przekonania:

1. Mózg to wielopoziomowy, hierarchiczny model generujący predykcje (tzw. model generatywny).
2. Hierarchicznej strukturze modelu odpowiada hierarchiczna anatomia kory mózgowej.
3. Celem działania modelu jest minimalizowanie błędów predykcyjnych, które wynikają z różnicy pomiędzy tworzonymi przez model przewidywaniami (predykcjami) a informacjami ze środowiska docierającymi przez zmysły.
4. Model jest hierarchiczny, ponieważ obejmuje wiele poziomów przetwarzania informacji.
5. Informacje w modelu są przetwarzane w dwóch kierunkach: odgórnie (predykcje dotyczące informacji docierającej do modelu) oraz oddolnie (informacje o potencjalnych błędach predykcyjnych). Oznacza to, że każdy poziom modelu przetwarzający informacje, generuje predykcje dotyczące tego, co się dzieje na poziomie znajdującym się poniżej, jednocześnie otrzymując z tego poziomu informacje dotyczące wielkości błędu predykcyjnego.

[1] Na temat koncepcji kompresji danych z wykorzystaniem kodowania predykcyjnego por. Shi, Sun, 1999.

6. Model jest skuteczny w działaniu, jeżeli predykcje generowane na wyższym poziomie prowadzą do minimalizacji błędów predykcyjnych pojawiających się na poziomie niższym.

1.1.1. Predykcja

Kluczowe w niniejszej koncepcji są pojęcia predykcji i modelu generatywnego. Pierwsze z nich można rozumieć na cztery podstawowe sposoby: (1) przewidywanie przyszłych stanów rzeczy; (2) przewidywanie przyszłych realizacji lub cech statystycznych zjawisk losowych; (3) korelacja, czyli statystyczny związek, zachodząca pomiędzy wartościami różnych zmiennych (np. zależność pomiędzy długością nauki w szkole a samooceną; albo zależność pomiędzy sygnałami odbieranymi przez gałki oczne – sygnały z lewego oka pozwalają skorelować sygnały z oka prawego); (4) wnioskowanie i testowanie hipotez. Zdaniem Anthony'ego Chemero i Michaela Andersona (2013, s. 204–205) w PP ma się do czynienia z predykcjami jako korelacjami i jako wnioskowaniami, przy czym, co wspomniani autorzy podkreślają, predykcja w sensie korelacji nie zakłada istnienia modelu i sieci powiązań pomiędzy informacjami (wiedza), zaś predykcje w sensie wnioskowania zakładają taki model (Clark, 2013a, s. 236). Te ostatnie są szczególnie istotne w PP, ponieważ to takie predykcje, jakie na podstawie parametrów wewnętrznego modelu przewidują strukturę przyczynową świata. Za ich pomocą model generatywny rekapituluje, tzn. syntetyzuje w pewnej postaci (jako surogat) statystyczną strukturę świata (Clark, 2013b, s. 182).

W niniejszej pracy, o ile nie zaznaczę inaczej, przez predykcje będę rozumiał statystyczne przewidywanie informacji pochodzących z sygnału zmysłowego, wytwarzane przez dany organizm, *resp.* podmiot, w sposób wielopoziomowy[2]. Predykcje niekoniecznie implikują przewidywanie przyszłych zjawisk, ponieważ jedna zmienna może przewidywać inną, o ile ta pierwsza przenosi informację o drugiej (mówimy wtedy o korelacji). Predykcje niskopoziomowe (np. dotyczące kształtu spostrzeganego obiektu) są szczegółowsze, odnoszą się do mniejszych wektorów czasoprzestrzennych, są krótkoterminowe. Predykcje wysokopoziomowe (np. dotyczące zachowania innych podmiotów czy podejmowania decyzji w warunkach ryzyka) mają charakter ogólniejszy, są bardziej abstrakcyjne i długoterminowe. Predykcją zatem będzie przykładowo przewidywanie, że poruszający się obiekt z dużym prawdopodobieństwem zmierza w moją stronę; jak i przewidywanie, że w trudnych warunkach atmosferycznych kierowcy samochodów zazwyczaj jeżdżą

[2] W sensie logicznym predykcje można uznać za rodzaj sądów dotyczących przyszłości (np. w formie zdania „Z pewnością będzie padać") albo sądów hipotetycznych (np. „Jeżeli założę palto, to prawdopodobnie nie zmoknę").

ostrożniej i zachowawczo. Należy dodać, że im bardziej abstrakcyjna i ogólna predykcja, tym bardziej kształtowana jest ona przez czynniki wewnętrzne (np. przekonania podmiotu, parametry modelu generatywnego, heurystyki), jak również czynniki zewnętrzne (ograniczenia środowiskowe itd.)[3]. Nowe predykcje powstają na bazie starych i są uzupełniane o informacje dotyczące błędu predykcyjnego. Tak rozumiane predykcje służą minimalizacji błędów predykcyjnych na każdym poziomie przetwarzania informacji przez organizm.

Przez błędy predykcyjne zaś należy rozumieć możliwe dysproporcje, które pojawiają się pomiędzy hipotezami tworzonymi na podstawie modelu świata posiadanego przez dany system poznawczy a informacjami pochodzącymi ze świata (określone zbiory danych), które są dostarczane przez wejścia zmysłowe. Innymi słowy, błąd predykcyjny to stopień niedopasowania pomiędzy dwoma sygnałami (gdzie większe niedopasowanie oznacza większy błąd). Można stwierdzić, że każda rozbieżność pomiędzy predykcją a sygnałem zmysłowym powoduje pojawienie się komunikatów o błędach predykcyjnych. W tym sensie błąd predykcyjny to różnica pomiędzy informacją zmysłową a predykcją (Hohwy, 2020a, s. 2).

1.1.2. Model generatywny

Pojęcie modelu generatywnego w kognitywistyce i neuronaukach nie jest nowe[4]. Niemniej żadne z jego wcześniejszych zastosowań nie miało tak wielu implikacji filozoficznych, jak ma to miejsce w przypadku jego użycia w PP. Modele generatywne można rozumieć jako spójne struktury zdolne do generowania szeregu zjawisk w sposób, który modeluje rzeczywisty proces, za pomocą którego zjawiska te są generowane. W sensie ogólnym modele generatywne są modelami statystycznymi, przez które należy rozumieć takie modele matematyczne, jakie uchwytują zależności zachodzące pomiędzy wartościami zbioru zmiennych losowych. Zależności te są reprezentowane przez parametry modelu, a zmienne mogą być obserwowane (tzn. niektóre z ich wartości mogą być dostarczane bezpośrednio przez dane) lub ukryte. W PP model generatywny jest rozumiany jako model statystyczny, którego zmienne są ukryte, tzn. nie da się jedno-jednoznacznie przyporządkować odpowiednich danych do wytwarzających ich procesów przyczynowych[5]. Modele statystyczne są na tyle dokładne, na ile dokładne są generowane przez nie predykcje.

[3] Szczegółowo omawiam te kwestie w rozdziale *Mechanizmy normatywne a działania w przetwarzaniu predykcyjnym*.

[4] Na temat jego rozumienia por. Ng, Jordan, 2001; Pickering, Clark, 2014.

[5] W tym sensie model generatywny to taki, jaki określa wspólne prawdopodobieństwo dla zbioru zmiennych, w przeciwieństwie do tzw. modelu dyskryminacyjnego, który określa warunkowy rozkład prawdopodobieństwa na zmienną docelową (często reprezentującą etykietę klasy) uwarunkowaną wartością (wartościami) obserwowanej (obserwowanych) zmiennej (zmiennych). Modele dyskryminujące są tak nazywane, ponieważ są zwykle (i naturalnie) wykorzystywane do klasyfikacji przedmiotów (por. Ng, Jordan, 2001).

Te zaś są oparte na określonych parametrach modelu, których model uczy się na podstawie odbieranych danych. Informacje i predykcje generowane przez model można opisać w kategoriach rozkładu prawdopodobieństwa. Z tego też powodu modele generatywne określa się jako „probabilistyczne". Model generatywny jest statystycznym modelem wspólnego rozkładu prawdopodobieństwa (*joint probability distribution*), który jakimś zmiennym losowym *X*, *Y*, ... przypisuje takie prawdopodobieństwo, że zmienne *X*, *Y*, … przyjmują określone wartości. Wspólny rozkład prawdopodobieństwa może również być wyrażony jako wspólna funkcja gęstości prawdopodobieństwa (w przypadku zmiennych ciągłych) lub wspólna funkcja masy prawdopodobieństwa (w przypadku zmiennych dyskretnych)[6]. Funkcje gęstości prawdopodobieństwa mogą potem być wykorzystane do znalezienia dwóch innych rodzajów rozkładów: rozkładu brzegowego (czyli rozkładu prawdopodobieństwa dla jakiejś jednej zmiennej losowej bez odniesienia do pozostałych zmiennych) oraz rozkładu warunkowego (czyli rozkładu prawdopodobieństwa dla wybranego podzbioru zmiennych losowych przy ustalonych wartościach pozostałych zmiennych). Badacze PP twierdzą, że model generatywny jest bayesowskim modelem probabilistycznym, co oznacza, że w jakiś sposób implementuje on regułę Bayesa[7]. Mówiąc o bayesowskim charakterze modelu generatywnego, ma się na myśli, że generowane przezeń predykcje oparte są na wcześniejszej wiedzy (*priors*) i jako takie są zakodowane w wewnętrznych parametrach modelu, a następnie są porównywane z przychodzącymi danymi i wykorzystywane do aktualizacji parametrów, tym samym określając nowe rozkłady prawdopodobieństwa określonych wartości zmiennych.

Dobrym przykładem modeli generatywnych są programy komputerowe, generujące realistyczne, dwu- lub trzywymiarowe obrazy ludzkich twarzy, opierając się jedynie na kilku zmiennych wejściowych, które odnoszą się do określonych parametrów, np. położenie mięśni twarzy, osadzenie oczu czy tekstura skóry (Kiefer, Hohwy, 2017, s. 3). Program ten realizuje proces generatywny, który wytwarza dane zjawiska, znajdując się w określony sposób w relacji do innego procesu generatywnego (np. takiego, za pomocą którego zmienne bazowe w świecie rzeczywistym przyczyniają się do pojawienia się rzeczywistych ludzkich twarzy). Stopień, w jakim proces generatywny może modelować inny proces czy zjawisko, zależy od podobieństwa zachodzącego pomiędzy tymi dwoma procesami. Inni badacze (por. Nair, Susskind, Hinton, 2008) dodają, że model generatywny to taki model, jaki wyjaśnia badane zjawiska poprzez określenie zbioru przyczyn, które mogły je wywołać. Oba te pojęcia modelu generatywnego – czyli (1) przyczynowo skutecznej struktury wytwarzającej zjawiska oraz (2) modelu, który wyjaśnia poprzez odniesienie do przyczyn danych zjawisk – są ze sobą ściśle powiązane.

[6] Można powiedzieć, że model generatywny jest statystycznym modelem tego, w jaki sposób generowane są obserwacje.

[7] Por. na ten temat rozdział *Przetwarzanie predykcyjne jako bayesowski model eksplanacyjny*.

Z teoretycznego punktu widzenia modele generatywne opisują tylko pewną zależność probabilistyczną, a to, czy interpretujemy ją jako przyczynową, czy np. informacyjną, zależy od kontekstu, w którym są stosowane. Upraszczając, możemy powiedzieć, że model jest taką strukturą, jaka jest w stanie wytworzyć dane zjawisko, o ile oferuje ona najlepsze wyjaśnienie tego zjawiska (por. Kiefer, 2017; Kiefer, Hohwy, 2018, s. 3; Seth, 2015).

Kiedy mówi się o hierarchicznym modelu generatywnym, to ma się zwykle na myśli hierarchie generatywnych modeli określonych i wzajemnie powiązanych ze sobą zjawisk. Modele hierarchiczne pozwalają wyjaśnić złożoność określonych zjawisk i wiążących je ze sobą zależności przyczynowych. Oznacza to, że wraz ze wzrostem złożoności modelu zwiększają się szanse na wygenerowanie odpowiedniego zjawiska z wykorzystaniem strukturalnie odmiennych zbiorów przyczyn.

Model generatywny odwzorowuje strukturę statystyczną pewnego zestawu obserwowanych danych wejściowych poprzez śledzenie (w efekcie schematycznego podsumowania – rekapitulację) macierzy przyczynowej odpowiedzialnej za tę strukturę. Dynamika jednostek kodujących taki model służy do przewidywania danych wejściowych do systemu. W ten sposób model generatywny generuje przyczyny na podstawie ich skutków, tj. treści i struktury sygnału zmysłowego. Model generatywny nieustannie i w sposób odgórny generuje predykcje, które przewidują treść i charakter sygnałów zmysłowych generowanych przez przyczyny zewnętrzne względem modelu. Proces ten ma na celu minimalizację pojawiających się błędów predykcyjnych rozumianych jako stopień niedopasowania pomiędzy dwoma sygnałami (por. Gładziejewski, 2016, s. 562).

Na koniec niniejszych analiz należy wspomnieć o pewnej istotnej dwuznaczności, która będzie co jakiś czas ujawniała się czytelnikowi i którą należy teraz wyjaśnić. Mianowicie: mówiąc o modelowaniu i modelach, można mieć na myśli dwie rzeczy: (1) czynność modelowania i tworzenia modeli, którą wykonują naukowcy; (2) oraz sam model generatywny jako organ poznawczy (Hohwy, 2015a), który modeluje czy reprezentuje odpowiednie zjawiska. W PP mówiąc o modelu, ma się przede wszystkim na myśli to, że podmiot bądź organizm ucieleśnia model swojego środowiska (który może być np. zlokalizowany w mózgu, jak twierdzą zwolennicy konserwatywnego PP, albo być organizmem jako takim, za czym argumentują zwolennicy radykalnych odmian PP)[8], nie zaś to, że mózg jest modelowany w taki, a nie inny sposób. Oczywiście istnieją odpowiednie algorytmy, które pozwalają modelować mózg jako probabilistyczny model generatywny, ale mówiąc o modelu, mam przede wszystkim na myśli jego rozumienie jako narządu, za pomocą którego organizm poznaje świat i weń ingeruje. Zatem kiedy będę używał pojęcia modelu generatywnego, to będę miał na myśli pewien narząd lub przedmiot, nie zaś wytwór badaczy, jakim jest np. model obliczeniowy albo model

[8] Por. na ten temat podrozdział *Konserwatywne przetwarzanie predykcyjne*.

geometryczny danego zjawiska. Pojęcie modelu ma zatem znaczenie przedmiotowe i metaprzedmiotowe. W niniejszej pracy, o ile nie zaznaczę inaczej, używam pojęcia modelu w sensie przedmiotowym.

Po tych koniecznych uwagach pojęciowych i definicyjnych mogę przejść do wstępnej charakterystyki PP. W najogólniejszym ujęciu można powiedzieć, że w myśl idei PP podstawowym zadaniem mózgu, rozumianym jako wielopoziomowy, hierarchiczny model generatywny, jest minimalizacja błędów predykcyjnych[9]. Minimalizacja błędów predykcyjnych ma kluczowe znaczenie dla organizmu[10], bowiem – w myśl głoszonego tu poglądu – wszelka percepcja jest podporządkowana celowi, jakim jest skuteczne działanie danego organizmu w środowisku[11]. Minimalizacja oznacza zredukowanie dysproporcji bądź różnicy pomiędzy określoną informacją zmysłową pochodzącą z sygnału zmysłowego a hipotezami dotyczącymi tego sygnału odniesionymi do wewnętrznych parametrów modelu. Model generatywny, *resp.* mózg, w celu minimalizacji potencjalnych błędów predykcyjnych w sposób ciągły tworzy statystyczne przewidywania (tj. predykcje) dotyczące tego, co wydarza się lub może wydarzyć się w świecie. Owe predykcje odnoszą się do obecnego i przyszłego kształtu informacji docierających do mózgu za pośrednictwem modalności zmysłowych.

1.1.3. Precyzja

Przetwarzanie predykcyjne (PP) oblicza błędy predykcyjne w ten sposób, że są one ważone za pomocą precyzji. Zatem ich skuteczne minimalizowanie zakłada określenie stopnia ich precyzji. Szacowanie bądź ważenie precyzji (*precision-weighted*) pozwala na stwierdzenie, na ile dany błąd jest precyzyjny, tzn. czy informacja, jaką on przekazuje, jest wiarygodna dla systemu, czy nie. Im większą precyzję system przypisuje danemu błędowi predykcyjnemu, tym mniejszy jest ten błąd (Friston, 2010). Jeżeli przykładowo spodziewam się, że dany błąd predykcyjny jest szczególnie wiarygodny czy wysoce prawdopodobny, to zwiększam jego wagę,

[9] Nie wchodzę tutaj w problem usytuowania modelu generatywnego. Część autorów twierdzi, że model ten jest zlokalizowany w mózgu (konserwatywne PP), inni bronią przekonania, że sam organizm jako taki jest modelem generatywnym (radykalne PP).

[10] Pojęć „organizm" i „system poznawczy" będę używał zamiennie, zawężając tym samym pojęcie systemów poznawczych do organizmów żywych. W innych partiach książki wprowadzam również pojęcie podmiotu, przez który rozumiem taki system poznawczy, jaki jest osadzony, ucieleśniony i aktywny w swoim środowisku. W anglojęzycznej literaturze poświęconej PP używa się pojęcia *agent*, które najlepiej byłoby tłumaczyć w języku polskim jego kalką, jaką jest *agent*, czyli ten, który działa. Niemniej, ze względu na słabe osadzenie filozoficzne tego pojęcia w polskiej literaturze oraz inne skojarzenia semantyczne, będę posługiwał się pojęciem podmiotu. Należy jednak to pojęcie rozumieć w sensie empirycznym, a nie transcendentalnym.

[11] Trzeba dodać, że minimalizacja w wielu przypadkach odbywa się w sposób zupełnie nieświadomy. Z tego też powodu należy powiedzieć, że treść tzw. doświadczenia fenomenalnego stanowi tylko część tego procesu (Wiese, Metzinger, 2017, s. 3).

czyli stopień, w jakim może on wpłynąć na parametry modelu. Precyzja jest zatem miarą stopnia niepewności[12].

Wyobraźmy sobie, że idąc lasem, dostrzegam dym. Nie jestem jednak pewny, czy jego przyczyną jest pożar lasu, czy ognisko. Wiem, że w tym czasie wiele osób pali ogniska w miejscach do tego przeznaczonych, niemniej dymu jest bardzo dużo i jego źródło stale się powiększa (o czym wnioskuję na podstawie pojawiającego się dymu). Przyjmuję hipotezę, zgodnie z którą prawdopodobniejsze jest to, że źródłem dymu jest pożar lasu (ponieważ o tej porze roku jest sucho i łatwo o pożar) niżeli wiele rozpalonych obok siebie ognisk (co wydaje się mało prawdopodobne). Z tego też powodu błąd predykcyjny dotyczący pożaru lasu, jako przyczyny pojawienia się dymu, jest dużo precyzyjniejszy niż błąd predykcyjny zakładający, że wiele ognisk jest przyczyną dymu. Zatem większą wagę przypisuję temu pierwszemu błędowi. W efekcie taki błąd ma mniejszy wpływ na parametry mojego modelu świata niż błąd odnoszący się do wielu ognisk.

Precyzja w PP jest rozumiana tak jak w statystyce, czyli jako wariancja odwrotna (*inverse variance*)[13]. Używane w PP takie pojęcia jak precyzja, niepewność czy zaskoczenie odnoszą się do własności rozkładów prawdopodobieństwa. Rozkłady prawdopodobieństwa są kluczowe, ponieważ za ich pomocą można sprawnie wytłumaczyć, na czym polega percepcja jako wnioskowanie. Mianowicie: proces wnioskowania wymaga reprezentowania rozkładów prawdopodobieństwa w odniesieniu do możliwych przyczyn pobudzeń zmysłowych. Te rozkłady pozwalają na oszacowanie wielu wartości. Przykładowo szeroki rozkład koduje wysoki stopień niepewności w odniesieniu do określonej przyczyny. Ta niepewność wiąże się ze stopniem zaskoczenia dotyczącym możliwych przyczyn przychodzącej informacji. Kluczową miarą niepewności jest szerokość lub wariancja rozkładu oraz jego odwrotność, czyli precyzja[14]. Precyzję można zatem określić również jako pewność dotyczącą danego przekonania (Adams, Brown, Friston, 2014, s. 1)[15].

---

[12] Pojęcie precyzji należy wiązać z tradycyjnie rozumianą uwagą. Trzeba jednak podkreślić, że odnosi się ona raczej do szacowania niepewności niż do tego, co jest reprezentowane (Feldman, Friston, 2010). Wysoka uwaga oznacza, że percepcja opiera się bardziej na oddolnie przychodzących informacjach pochodzących z sygnału zmysłowego, ponieważ każde małe odchylenie jest szybko zauważane. Niski stopień uwagi oznacza zaś, że w danej sytuacji percepcja w większym stopniu opiera się na odgórnie generowanych predykcjach, np. postrzegany jest tylko niejasny zarys danego przedmiotu.

[13] Wiąże się to przede wszystkim z faktem, że PP uważa się za modelowania bayesowskie. Zagadnieniem tym zajmuję się szczegółowo w rozdziale *Przetwarzanie predykcyjne jako bayesowski model eksplanacyjny*.

[14] Kluczowy dla statystyki i PP jest rozkład normalny, zwany również „rozkładem Gaussa". Wykres funkcji prawdopodobieństwa tego rozkładu jest krzywą w kształcie dzwonu. Rozkład ten jest bardzo ważny ze względu na jego częstość występowania w przyrodzie. Jest ciągłym rozkładem prawdopodobieństwa z maksimum dla wartości średniej. Wartość średnia określa położenie centralnego maksimum krzywej. Szerokość krzywej określa tzw. wariancja. Opisuje ona „rozrzut" danej wartości od wartości średniej. Innymi słowy, im większa jest wariancja, tym mniejsza precyzja, co oznacza, że typowe odchylenie wyniku od średniej jest większe. Tak więc mniejsza wariancja oznaczać będzie większą precyzję, bo typowe odchylenie będzie mniejsze, a tym samym rozkład bardziej zwarty w sobie.

[15] Przekonanie zaś w PP należy rozumieć jako rozkład prawdopodobieństwa dotyczący nieznanych stanów rzeczy lub własności świata. Innymi słowy, przekonanie to systemowe przekonanie (*prior*) o wysokim stopniu abstrakcji, tj. odnoszące się do ogólnej wiedzy o świecie.

Z precyzją wiąże się dokładność, która oznacza stopień zbliżenia pomiaru do rzeczywistej wartości wielkości mierzonej. Precyzja zaś oznacza stopień, w jakim powtarzane pomiary w niezmienionych warunkach dają takie same wyniki. Innymi słowy, precyzja to możliwość odnotowania różnic. Oznacza to, że czym innym jest precyzja dotycząca błędu predykcyjnego, a czym innym jego dokładność[16]. Z szacowaniem precyzji wiąże się zatem szum (dla wielkości nielosowych) oraz błąd (dla wielkości losowych).

1.1.4. Hierarchiczność

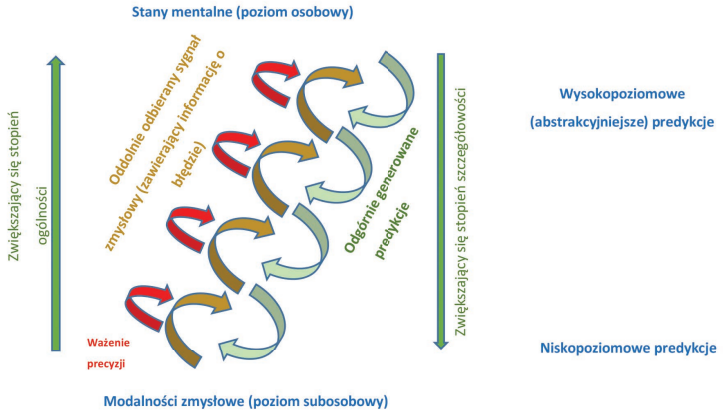
Po tym omówieniu musimy wrócić do struktury modelu. Predykcje minimalizujące błędy są hierarchicznie zorganizowane i generowane na poszczególnych poziomach modelu generatywnego. Dzięki temu przewidywania obecne na wyższych poziomach modelu odnoszą się do informacji o błędach predykcyjnych, które są obecne na niższych poziomach. Ściślej mówiąc, predykcje w sposób odgórny organizują oddolnie przychodzące i pochodzące ze zmysłów informacje w celu możliwie dużego zminimalizowania zaskoczenia (*surprisal*), jakie wiąże się z nieznajomością przyczyn pobudzeń zmysłowych. Innymi słowy, w dół hierarchii wysyłane są predykcje, w górę zaś informacje dotyczące rozmiaru błędu predykcyjnego. Informacje te zaś są precyzyjnie szacowane w celu określenia ich wiarygodności.

Powyższa struktura modelu wiąże się z obserwacją, według której mózg jest zorganizowany hierarchicznie. Organizacja ta została najlepiej zbadana w odniesieniu do systemu wzrokowego. Kolejne obszary korowe można uznać za tworzące hierarchię. Niższe obszary (poziomy) tej hierarchii wiążą się z wejściami zmysłowymi, wyższe zaś z multimodalnością i funkcjami asocjacyjnymi. Pojęcie hierarchii wiąże się przede wszystkim z tzw. połączeniami zstępującymi (*forward*) i wstępującymi (*backward*), które oparte są na specyfice warstw korowych, odpowiedzialnych za wewnętrzne połączenia w mózgu (por. Felleman, Van Essen, 1991; Rao, Ballard, 1999). W ujęciu PP korowe hierarchie, realizując połączenia zstępujące, generują predykcje, które odnoszą się do informacji na temat przyczynowej struktury określonych stanów rzeczy, które są zawarte w połączeniach wstępujących i kodowane pod postacią rozkładów prawdopodobieństwa. Połączenia wstępujące zaś przekazują informację zwrotną, dotyczącą błędu predykcyjnego na wyższe poziomy. Połączenia „góra – dół" i „dół – góra" są zatem połączeniami zwrotnymi i są one realizowane na każdym poziomie hierarchii (Friston, Stephan, 2007, s. 443).

Model generatywny budowany jest w odniesieniu do wektorów czasoprzestrzennych. Dwukierunkowa (w sensie przetwarzania informacji), hierarchiczna struktura

[16] Przekonująco uzasadnia to Johan Kwisthout, Harold Bekkering, Iris van Rooij (2017), twierdząc, że dla szacowania precyzji wcale nie jest konieczne, aby informacja o błędzie była bardzo szczegółowa.

modelu umożliwia przetwarzanie określonej informacji jedynie w odniesieniu do konkretnych skal czasowych i zrelatywizowanych do nich wymiarów przestrzennych. Jest ona ograniczona do różnicy pomiędzy wyższym a niższym poziomem modelu (por. Friston, 2003). Mianowicie: predykcje na (przykładowym) poziomie (oznaczmy go jako *x*) minimalizują błąd predykcyjny na poziomie *x–1*, ale nie mogą tego robić na poziomie *x–2*. Za to predykcje na poziomie *x–1* minimalizują błąd na poziomie *x–2* itd. (por. Friston, Kiebel, 2009; Hohwy, 2015a). Przyczyną takiego stanu rzeczy są różnice czasowe pomiędzy poziomami (por. Williams, 2018)[17]. Czym wyżej posuwamy się w hierarchii, tym większych wymiarów czasowych dotyczą kolejne poziomy modelu. Jednocześnie zwiększa się poziom abstrakcji poszczególnych poziomów: od mało abstrakcyjnych, np. dotyczących detekcji krawędzi, do bardziej abstrakcyjnych, odwołujących się przykładowo do procesów kategoryzacji. Można zatem powiedzieć, że kolejne poziomy modelu są budowane na podstawie jego czasoprzestrzennej struktury, a nie określonych własności środowiska. Innymi słowy, nie są one odróżniane ze względu na to, co reprezentują, czyli ze względu na treść, ale ze względu na skalę czasu, w której funkcjonują (Hohwy, 2013, s. 72).



Rysunek 1. Schematyczne przedstawienie poziomów przetwarzania informacji przez model generatywny. Oddolnie przychodzące informacje zmysłowe są przetwarzane w kontekście wiedzy i przekonań (*priors*) podmiotu (tj. wewnętrznych parametrów modelu generatywnego), pochodzących z wyższych poziomów modelu generatywnego. Nieprzewidziane informacje pochodzące z sygnału zmysłowego są przenoszone w górę hierarchii, co prowadzi do wygenerowania i dostosowania odpowiednich predykcji. Informacja o błędzie jest ważona za pomocą precyzji, a jego „pozostałości" (*residuals*) przenoszone są na wyższe poziomy, gdzie następuje ich minimalizacja. Cały proces jest ciągły i powtarzalny (opracowano na podstawie: Clark, 2016, s. 30, 59)

[17] Daniel Williams z tego powodu krytykuje hierarchiczność w PP jako nierealistyczną.

1.2. Możliwe algorytmy przetwarzania predykcyjnego i filozoficzne inspiracje

Przetwarzanie predykcyjne (PP) jest nie tylko nietrywialną koncepcją tego, w jaki sposób realizowane jest ludzkie poznanie, ale także ujęciem, które stara się wyjaśnić, w jaki sposób mózg wykonuje wnioskowania probabilistyczne. Odwołanie się do trzech poziomów analizy wyróżnionych przez Marra (1982) pozwoli dobrze wyjaśnić, na czym polega PP i do czego się odnosi. Zdaniem Marra systemy poznawcze można analizować na poziomie obliczeniowym, algorytmicznym oraz implementacji. Poziom obliczeniowy odpowiada na pytanie o to, *co* robi dany system i *dlaczego*. Chodzi zatem o opis zachowania danego systemu. Poziom algorytmiczny pozwala odpowiedzieć na pytanie dotyczące tego, w *jaki* sposób dany system (mechanizm) robi to, co robi, jednak nie określa, jak i gdzie fizycznie realizowany jest dany system. To jest zadanie teorii budowanych na poziomie implementacji, który wskazuje na miejsce w mózgu lub innym nośniku, *gdzie* dany system jest fizycznie bądź biologicznie realizowany i *jak* to się dzieje[18].

Można zatem powiedzieć, że analiza na poziomie obliczeniowym wykazuje, że mózg jest zaangażowany w tworzenie modelu generatywnego środowiska na drodze realizacji jakiegoś wnioskowania probabilistycznego. Wiedząc to, można zapytać, na mocy jakich reguł mózg może być probabilistycznym modelem generatywnym? W dotychczasowej literaturze można wymienić wiele możliwych algorytmów, które mają odpowiedzieć na powyższe pytanie. Michael Spratling (2017) wymienia pięć teorii algorytmicznych, które można opisać mianem PP. Są to:

- liniowe kodowanie predykcyjne (por. Makhoul, 1975);
- kodowanie predykcyjne w siatkówce oka (por. Srinivasan, Laughlin, Dubs, 1982);
- kodowanie predykcyjne w pierwotnej korze mózgowej oparte na modelu Rajeshy Rao i Dana Ballarda (1999);
- kodowanie predykcyjne w pierwotnej korze mózgowej oparte na modelu BC (*biased comptetition*);
- kodowanie predykcyjne oparte na modelu FEP (*free-energy principle*)[19].

Wszystkie wymienione algorytmy mają na celu dopasowanie modelu generatywnego do danych sensorycznych. Różnią się one jednak ujęciami modelu generatywnego, który wykorzystują, oraz sposobem, w jaki są one powiązane z neurobiologią. Na potrzeby niniejszych analiz omówię algorytmy Rao i Ballarda oraz kodowanie liniowe, ponieważ wydaje się, że są one najistotniejsze z punktu widzenia późniejszych analiz PP. Algorytm oparty na FEP zostanie przeze mnie

[18] Propozycja Marra jest szeroko analizowana przeze mnie w rozdziale *Przetwarzanie predykcyjne jako bayesowski model eksplanacyjny*.

[19] Argumenty przeciwko traktowaniu FEP jako możliwego algorytmu dla PP omawiam w rozdziałach *Zasada energii swobodnej w przetwarzaniu predykcyjnym* i *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego*.

szczegółowo przeanalizowany w rozdziale *Zasada energii swobodnej w przetwarzaniu predykcyjnym*.

W 1999 roku Rao i Ballard opublikowali artykuł pt. *Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects*. Wydaje się, że tę pracę można uznać za właściwy impuls dla rozwijanych obecnie w kognitywistyce różnych ujęć PP[20]. Rao i Ballard zaproponowali model przetwarzania informacji wzrokowej, w którym informacje przekazywane z wyższych do niższych warstw kory mózgowej zawierają przewidywania dotyczące aktywności neuronów obecnych na jej niższych warstwach. Zależność poszczególnych warstw w tym modelu można rozumieć w kategoriach sprzężenia zwrotnego. Połączenie zstępujące (odgórne) niesie informacje o przewidywaniach dotyczących aktywności neuronalnej niższych warstw korowych. Oddolnie zaś generowane są informacje o możliwych błędach zachodzących pomiędzy odgórnymi przewidywaniami a faktyczną aktywnością neuronalną. Informacja o tych błędach jest następnie wykorzystywana przez mózg do tworzenia aktualnych przewidywań dotyczących charakteru sygnału odbieranego przez wejścia zmysłowe. Na podstawie dokonanej korekty, tworzone są nowe przewidywania (Rao, Ballard, 1999, s. 80). Rao i Ballard podkreślają hierarchiczny i odgórny charakter modelu wykorzystującego kodowanie predykcyjne. Niskopoziomowe przewidywania dotyczą czasoprzestrzennych parametrów przychodzących informacji. Przewidywania generowane na wyższych poziomach modelu mają zaś abstrakcyjniejszy i ogólniejszy charakter. Pomysł Rao i Ballarda został wykorzystany przez innych badaczy i do dziś jest to hipoteza płodna poznawczo w obrębie neuronauk.

Niektórzy badacze twierdzą (por. Orlandi, 2017), że istotne dla obecnie rozwijanych ujęć PP są nie tylko badania z zakresu neuronauk, ale także inspiracje pochodzące od badaczy transmisji obrazów w telewizji (m.in. Harrison, 1952; Kretzmer, 1952; Oliver, 1952). Konieczna dla przekazu telewizyjnego jest ciągła transmisja obrazów. Istnieje jednak możliwość przeciążania odpowiednich przekaźników. W jaki zatem sposób powiązać konieczność nieprzerwanego przekazu z ograniczoną mocą medium, które ten przekaz będzie realizować? Rozwiązaniem miało być wykorzystanie tzw. liniowego przewidywania (*linear prediction*). Załóżmy, że chcemy transmitować „nieruchomy" obraz. Istnieją dwie możliwości: albo transmituje się ten sygnał dwa i więcej razy, ryzykując przeciążenie, albo transmituje się nieruchomy sygnał raz, a następnie przekazuje się informację dotyczącą możliwych zmian w obrazie z poprzedniej transmisji (Harrison, 1952, s. 766). Analogicznie, transmitując zmieniający się obrazek, możemy mieć sygnał, który informuje o każdym aspekcie zmieniającego się obrazu, przeciążając tym samym przekaz, albo transmituje się sygnał informujący tylko o tym, jakie zmia-

[20] Oprócz tej pracy można jeszcze wspomnieć o artykułach Tai Sing Lee i Davida Mumforda (2003) oraz Karla J. Fristona (2005). Por. Clark, 2013b, s. 182.

ny nastąpiły w stosunku do obrazu początkowego. Przesyłana jest zatem tylko informacja dotycząca różnicy pomiędzy sygnałami. Inaczej mówiąc, podczas transmisji sygnału oczekuje się, że nowy obraz będzie w przybliżeniu podobny do poprzedniego. Tym sposobem tworzone jest przewidywanie dotyczące tego, jakie powinny być następujące wartości pikseli w przychodzącym sygnale. Przewidywanie to jest następnie potwierdzane lub obalane przez przychodzący sygnał, co można rozumieć w kategoriach minimalizacji błędu (por. Orlandi, 2017, s. 2).

Dla filozoficznych inspiracji PP istotne są również prace takich badaczy jak Donald M. MacKay (1956), Ulrich Neisser (1967) czy Richard Gregory (1980), których zalicza się do grupy psychologów poznawczych będących zwolennikami „analizy przez syntezę" (*analysis-by-synthesis*). W myśl tego poglądu mózg nie buduje wewnętrznego modelu świata, gromadząc (oddolnie) odbierane przez zmysły informacje, ile raczej próbuje utworzyć takie reprezentacje, jakie następnie porównuje z tymi informacjami, wybierając te, które są z nimi zgodne.

Przykładowo, w pochodzącym z 1980 roku artykule *Perceptions as hypothesis* Gregory powiada, że percepcja opiera się na odgórnie realizowanych procesach przetwarzania informacji. Informacje pochodzące ze środowiska i docierające do mózgu za pośrednictwem wejść zmysłowych są niejednoznaczne i niedookreślone. Aby móc je zinterpretować, system poznawczy musi korzystać ze swoich wcześniejszych doświadczeń oraz ze zgromadzonej wiedzy. Za ich pomocą wnioskuje o tym, co jest spostrzegane. Percepcję – stwierdza Gregory – można zatem porównać do naukowego procesu stawiania hipotez. Zarówno hipotezy w nauce, jak i procesy percepcyjne pozwalają rozpoznawać odpowiednie sytuacje przez przedmioty w różnych, zmiennych kontekstach, na podstawie szczątkowych oraz niejasnych informacji. Z tego powodu można powiedzieć, że procesy percepcyjne są procesami tworzenia hipotez (Gregory, 1980, s. 182). Za hipotetycznym charakterem procesów percepcyjnych mają przemawiać następujące argumenty:

1. Tylko niewielka część z informacji, które trafiają do oka, jest przesyłana do mózgu (Gregory szacuje, że ok. 10%).
2. Mózg niejako musi zgadywać, co dana osoba spostrzega. Na tej podstawie tworzy percepcję rzeczywistości.
3. Bez założenia hipotetycznej natury percepcji trudno jest wyjaśnić istnienie złudzeń optycznych, takich jak sześcian Neckera[21]. Błędne hipotezy percepcyjne umożliwiają pojawienie się tego rodzaju iluzji wzrokowych.

[21] W 1832 roku Louis Necker opublikował rycinę, która przedstawiała sześcian zmieniający swoje położenie podczas oglądania. Ów efekt osiągnął, usuwając z obrazka wszelkie percepcyjne wskazówki dotyczące głębi. Patrząc na figurę przedstawioną na rycinie, widoczny jest tylko układ linii, niemniej spodziewamy się zobaczyć sześcian. Mózg zatem, w myśl koncepcji Gregory'ego, tworzy hipotezę, że spostrzegany przedmiot jest sześcianem.

4. Istnienie hipotez percepcyjnych pozwala wyjaśnić spostrzeganie tzw. przedmiotów niemożliwych, takich jak trójkąt Penrose'a czy wstęga Möbiusa[22].

Dla rozważań prezentowanych w niniejszej pracy równie istotne są filozoficzne wnioski wyłaniające się z analiz neuronaukowców. Pomimo że PP ugruntowane jest w badaniach empirycznych i anatomicznych, to leżące u jego podstaw założenia i inspiracje mają często charakter czysto filozoficzny. Twierdzę, że ramy wyznaczone przez neuronaukowe analizy kodowania predykcyjnego pozwalają na stworzenie nietrywialnej koncepcji wyjaśniającej funkcjonowanie systemu poznawczego oraz mechanizmów przezeń realizowanych. Jak niebawem wykażę, istnieje kilka możliwych wykładni tego podejścia i za każdą z nich można podać mocne argumenty. Omówienie tych stanowisk jest istotne dla uwypuklenia koncepcji, która będzie rozwijana na kartach tej książki.

Hohwy, czyli jeden z najważniejszych filozoficznych autorów zajmujących się PP, twierdzi, że źródeł tej koncepcji należy doszukiwać się w myśli arabskiego filozofa żyjącego na przełomie X i XI wieku, tj. Ibn al-Hajtama. Napisał on dzieło pt. *Optyka,* w którym proces widzenia przedmiotów tłumaczył odbijaniem się światła od ich powierzchni i skierowaniem tego światła do oczu. Bronił poglądu, według którego wiele widzialnych własności przedmiotów jest ujmowanych za pośrednictwem sądzenia i wnioskowania (inferencji) (Hohwy, 2013, s. 5).

Ideę PP wielu badaczy wiąże z pracami XIX-wiecznego neokantysty Hermanna von Helmholtza[23]. Był on pierwszym myślicielem, który wprost stwierdził, że mózg jest maszyną testującą hipotezy. Helmholtz (1867, s. 430) powiada, że „aktywności psychiczne", na których oparta jest percepcja, są raczej nieświadome niż świadome, ponieważ:

> na podstawie zaobserwowanego wpływu na nasze zmysły dochodzą do pojęcia przyczyny tego wpływu. Dzieje się tak pomimo tego, że w rzeczywistości zawsze mamy bezpośredni dostęp do zdarzeń odbieranych przez nasz układ nerwowy, co oznacza, że odczuwamy tylko skutki, nie mamy zaś bezpośredniego dostępu do przedmiotów zewnętrznych.

W innym miejscu dodaje (Helmholtz, 1995, s. 13):

> zmysły rzeczywiście dostarczają nam informacji o wytwarzanych na nich efektach zewnętrznych, ale przekazują te efekty do naszej świadomości w zupełnie innej formie. Oznacza to, że charakter zmysłowej percepcji zależy nie tyle od właściwości obiektu spostrzeganego, ile od własności organu zmysłowego, za pośrednictwem którego otrzymujemy informację.

[22] Figury niemożliwe są to takie przedstawienia trójwymiarowych figur na płaszczyźnie, dla jakich nie jest możliwe skonstruowanie ich trójwymiarowych odpowiedników. Najbardziej znanymi przedstawieniami zawierającymi takie figury są grafiki Mauritsa C. Eschera.

[23] Zahavi (2018) wskazuje na odpowiednie elementy koncepcji XIX- i XX-wiecznych neokantystów, obecne w rozważaniach przynajmniej niektórych zwolenników koncepcji przetwarzania predykcyjnego.

Powołując się na Helmholtza, powinno się zatem mówić o percepcji jako nieświadomym wnioskowaniu (Hohwy, 2013, s. 18). Takie ujęcie jednak budzi opór wielu autorów i samo w sobie jest bardzo niejasne. Co to znaczy, że percepcja jest nieświadomym wnioskowaniem? Czy wyjaśnienie nieświadomie realizowanych procesów w mózgu stanowi ostatecznie wyjaśnienie procesów percepcyjnych, które są świadome? Czy, ogólniej mówiąc, wyjaśnienie pracy mózgu może zastąpić myślenie o poznaniu w kategoriach umysłu, jego funkcji i architektury? Propozycja Helmholtza sugeruje, że granicą umysłu jest fizyczna granica mózgu, czyli czaszka (por. Gładziejewski, 2017b; Hohwy, 2014, s. 1[24]).

Wielu autorów przypisuje zwolennikom koncepcji PP powrót do XIX-wiecznych spekulacji dotyczących poznawalności rzeczywistości, różnicy pomiędzy zjawiskami a rzeczami samymi w sobie czy pytań o idealizm. W tym kontekście Zahavi (2018) zarzuca omawianym tutaj ujęciom neuroreprezentacjonizm, zgodnie z którym cała treść świadomego doświadczenia jest konstrukcją czy symulacją tworzoną przez mózg. Jest to mocny zarzut i prowadzi do ogólnego stwierdzenia, że inspirujące się Kantem i neokantyzmem ujęcia PP są w istocie neurocentryczne i dalekie od myślenia o poznaniu w kategoriach tzw. 4E, czyli tego ujęcia procesów poznawczych, które podkreśla ich ucieleśnienie (*embodied*), osadzenie w środowisku (*embedded*), rozszerzenie (*extended*) oraz działanie w nim (*enacted*). Dodatkowo Zahavi zarzuca zwolennikom PP niemożliwość uzasadnienia realizmu, czyli przekonania, zgodnie z którym świat przedmiotów danych w doświadczeniu obiektywnie, tzn. niezależnie od naszych władz poznawczych, istnieje. W jego opinii (1) zwolennicy PP mają tendencję do zakładania istnienia mózgu (i pewnych rodzajów reprezentacji w nim) w swoich wyjaśnieniach, ale (2) nie są do tego uprawnieni, ponieważ reprezentacjonizm pociąga za sobą sceptycyzm wobec reprezentowanego świata, w tym także mózgu. Innymi słowy, nie mamy możliwości zagwarantowania, że cokolwiek w prawdziwym świecie zewnętrznym odpowiada tym reprezentacjom, a także nie mamy gwarancji, że świat zewnętrzny („rzecz sama w sobie" Kanta) istnieje lub w jakikolwiek sposób jest dla nas zrozumiały. Z tego też powodu PP, zdaniem Zahaviego, prowadzi do antyrealizmu[25].

Zahavi w swojej krytyce idei PP nie jest odosobniony. Część badaczy krytykuje PP za jego radykalny neuroreprezentacjonizm, konstruktywizm, kantyzm czy nawet idealizm (por. Anderson, Chemero, 2013; Anderson, 2017; Orlandi, 2017). Nietrudno dziwić się takiego rodzaju krytykom, bowiem niektórzy zwolennicy PP twierdzą, że percepcja jest „kontrolowaną halucynacją" lub doświadczeniem „wirtualnej rzeczywistości" (Hohwy, 2013, s. 137–138). Wiąże się to z głoszonym

[24] Przetwarzanie predykcyjne (PP) „ujawnia, że umysł jest odizolowany od świata, bowiem wydaje się, że jest bardziej neurocentrycznie związany z czaszką niż ucieleśniony lub rozszerzony, zaś samo jego działanie jest bardziej inferencyjnym procesem na wejściu zmysłowym niż enaktywnym sprzężeniem ciała ze środowiskiem".

[25] Przez antyrealizm w najbardziej ogólnym sensie rozumiem tutaj pogląd, który odrzuca bądź zawiesza tezę realistyczną.

przez predykcjonizm przekonaniem, według którego z powodu nieznajomości przyczynowych źródeł pobudzeń zmysłowych mózg niejako musi je „zgadywać" (Clark, 2016, s. 2), czyli stawiać na ich temat hipotezy percepcyjne. Wielu z nich również podkreśla kantowski rodowód PP (Clark, 2013b; Gładziejewski, 2016, 2017a; Hohwy, 2013; Swanson, 2016). Twierdzi się tak m.in. z powodu odgórnego (*top-down*) sposobu tłumaczenia poznania, aktywnego i receptywnego charakteru systemu poznawczego czy podkreślania znaczenia wewnętrznych reprezentacji w procesie poznawania świata.

Krytyka Zahaviego i innych autorów wymaga odpowiedzi, ponieważ wydaje się, że przyjęcie ich argumentacji może prowadzić do konkluzji, iż mamy nie tylko do czynienia z koncepcją odnawiającą stare spory epistemologiczne, ale także skazaną na szereg metafizycznych aporii rodem z XIX wieku. Zanim jednak odpowiem na wspomniane krytyki, konieczne jest zastanowienie się nad tym, jaki jest filozoficzny rodowód PP i jego faktyczne zastosowanie. Wpierw zastanowię się nad tym, na ile PP można uznać za ujęcie postkantowskie (podrozdział *Kant i przetwarzanie predykcyjne*). Następnie pokażę, że w PP nie mamy do czynienia z uzasadnieniami transcendentalnymi, ale z tzw. wnioskowaniem abdukcyjnym (podrozdział *Kant i przetwarzanie predykcyjne*). Na koniec przejdę do omówienia krytyki PP przeprowadzonej przez Zahaviego (2018) oraz odpowiem na sformułowany przez niego zarzut (podrozdział *Krytyka przetwarzania predykcyjnego*). Tym samym wykażę, że PP nie stanowi prostego rozwinięcia XIX-wiecznych neokantowskich rozwiązań w filozofii i fizjologii, ale interesującą i nowatorską propozycję opisu oraz wyjaśnienia wielu zjawisk poznawczych.

1.3. Kant i przetwarzanie predykcyjne

Powiada się, że Kant pogodził empiryzm z aprioryzmem, głosząc postulat pochodzenia całej wiedzy z dwóch równorzędnych źródeł, tj. doświadczenia oraz czystej aktywności intelektu. Wskazał on na trzy podstawowe władze umysłu, czyli „podmiotowe źródła poznania, na których opiera się możliwość doświadczenia w ogóle i poznania jego przedmiotów: zmysł, wyobraźnia i apercepcja" (Kant, 2001, s. 144). Zmysły to władza pasywna, czyli podmiotowe zdolności do bycia pobudzonym i modyfikowanym. Ujmują one swoje przedmioty w sposób intuicyjny, czyli wprost. Wyobraźnia jest spontaniczna i syntetyzuje materiał dostarczony przez zmysły. Apercepcja zaś, czyli intelekt, to dyskursywna władza pojęć. Inaczej niż zmysłowość, nie prezentuje ona swoich przedmiotów bezpośrednio, lecz myśli o nich za pomocą pojęć. Jest spontaniczna i samorzutna. Współpraca tych trzech władz w odniesieniu do zjawisk umożliwia doświadczenie empiryczne (Maciejczak, 2007, s. 25 i nn.). Wielość wrażeń docierających do umysłu za pośrednictwem

zmysłów jest kategoryzowana i syntetyzowana przez tzw. transcendentalne formy naoczności (czas i przestrzeń) oraz przez formy intelektu. Dane naoczne muszą zostać ujęte w ramy transcendentalnych pojęć, aby mogły zostać przez podmiot rozpoznane i wykorzystane.

Z punktu widzenia poruszanych tutaj zagadnień istotna jest metoda badawcza, którą posługuje się Kant. Dokonany przez niego w *Krytyce czystego rozumu* tzw. przewrót kopernikański, zgodnie z którym przedmiot nie jest nigdy dany w sposób niezależny od podmiotu, ściśle wiązał się z problemem konstytucji, czyli koniecznością wyjaśnienia sposobu formowania się przedmiotu w procesie poznania. Proces konstytucji przedmiotu zakłada jego warstwową budowę. Owa budowa wiąże się po pierwsze z kantowskim podziałem na władze poznawcze, po drugie zaś z samą ideą przewrotu. Kolejnym warstwom przedmiotowym odpowiadają kolejne poziomy konstytucji, której podstawowym mechanizmem jest synteza (Maciejczak, 2007, s. 36). W spostrzeżeniu jako takim obcujemy już z przedmiotowością ukonstytuowaną. Z tego powodu pytanie o konstytucje ma charakter hipotetyczny i regresywny, bowiem „jest niejako wyszukiwaniem przyczyn do danego skutku" (Kant, 2001, s. 28). Metoda ta jest szczególnie widoczna w części *Krytyki* poświęconej dedukcji czystych pojęć intelektu. W kantowskiej dedukcji transcendentalnej nie chodzi o formalno-logiczne uzasadnienie jednych twierdzeń przy użyciu innych, ale o wyprowadzenie jednych zdań, czyli konkluzji, z innych zdań, tj. poprzedników. Tak rozumiana dedukcja objaśnia sposób, „w jaki pojęcia mogą *a priori* odnosić się do przedmiotów" (Kant, 2001, s. 124). Taka „analiza regresywna" (Höffe, 2003, s. 89), poprzez sprowadzenie kategorii do ich źródła, wykazuje, że bez kategorii nie jest możliwa konstytucja i poznanie przedmiotów danych w doświadczeniu.

Współcześnie takie regresywne postępowanie określa się mianem „inżynierii odwrotnej" (*reverse engineering*) (por. Dennett, 1995a). Inżynieria odwrotna to taki sposób postępowania, w jakim analizuje się gotowe urządzenie (przedmiot, strukturę, mechanizm) w celu poznania sposobu jego działania oraz funkcji, które realizuje. Nie chodzi zatem o utworzenie nowego urządzenia, ale o zrozumienie jego budowy, tj. odtworzenie jego konstrukcji. Można powiedzieć, że celem inżynierii odwrotnej jest dostarczenie hipotez na temat mechanizmów (transcendentalnych, psychologicznych czy biologicznych). Mechanizmy te zaś mają wyjaśniać dane zjawiska (Miłkowski, 2004). Współcześnie metodę inżynierii odwrotnej wiąże się z naturalizmem oraz ewolucjonizmem (por. Dennett, 1995a). Niektórzy jednak twierdzą, że takie podejście można przypisać Kantowi, który miałby stosować tę metodę w ramach dedukcji transcendentalnej (por. Swanson, 2016). Przemawia za tym następujący argument: funkcjonalna analiza charakterystyczna dla inżynierii odwrotnej ma charakter odgórny. Dzięki temu pozwala na rozpoznanie sposobu, w jaki części danego systemu osiągają swoje funkcje. Chodzi zatem o wskazanie

tych elementów bądź części danego mechanizmu, które są konieczne dla jego funkcjonowania, czyli dzięki którym system może działać tak, jak działa.

Istnieją jednak dwa silne argumenty odmawiające Kantowi stosowania metody inżynierii odwrotnej. Po pierwsze charakterystyczne dla transcendentalizmu jest abstrahowanie od wszelkiej maści pytań genetycznych, czyli tych, które dotyczą pochodzenia struktury poznawczej. Dla badania warunków możliwości doświadczenia nieistotne są zatem kwestie filogenetyczne, ontogenetyczne, historyczne czy kulturowe. Innymi słowy, nie ma tutaj miejsca na szukanie odpowiedzi na pytanie np. o to, dlaczego formy zmysłowości, takie jak czas i przestrzeń, mają taki, a nie inny charakter. Podobnie nie ma potrzeby pytać o to, dlaczego ludzie posiadają taki, a nie inny system kategorii, albo stosują takie, a nie inne pojęcia. Abstrahowanie od wyjaśnień funkcjonalnych jest drugim argumentem przemawiającym za tym, że Kantowska metoda nie była metodą inżynierii odwrotnej. Mianowicie: jeżeli przez funkcjonalne wyjaśnienie zjawiska (charakterystyczne dla inżynierii odwrotnej) rozumieć takie, jakie odwołuje się do funkcji, którą dane zjawisko spełnia, to metoda transcendentalna nie ma takiego charakteru. Wyjaśnienia na poziomie transcendentalnym mają inny charakter, nie odnoszą się bowiem do sfery faktów, ale do dziedziny, która ma fakty uzasadniać. Pytanie zatem nie dotyczy zastosowania danej struktury, ale tego, jak jest możliwe, że robi to, co robi, czyli np. jak umożliwia konstytucję przedmiotu danego w doświadczeniu (Leszczyński, 2010, s. 43–46). Z tego to powodu Kant odróżnił pytania, które dotyczą faktów (*quid facti*), od pytań, które odnoszą się do tego, jakim prawem coś się dzieje (*quid juris*) (Kant, 2001, s. 123). Te pierwsze pytania miałyby być typowe dla inżynierii odwrotnej, zaś pytania o uprawnienie faktów miałyby być charakterystyczne dla metody transcendentalnej[26].

Siła wyżej przedstawionych argumentów jest zależna od interpretacji propozycji Kanta. Problem ten jednak nie jest istotny dla rozważań prezentowanych w niniejszej pracy. Istotne jest to, czy i na ile podejście oparte na idei PP ma charakter kantowski. Jest to ważne, ponieważ wielu badaczy powołuje się na myśl królewieckiego filozofa, kiedy próbują wskazać źródła koncepcji predykcyjnej (por. Anderson, Chemero, 2013, s. 204; Clark, 2013b, s. 196; Gładziejewski, 2016, 2017a; Hohwy, 2013, s. 5; Swanson, 2016). Link R. Swanson wprost deklaruje, że PP bezpośrednio wywodzi się z transcendentalizmu Kanta. Przemawiać mają za tym następujące fakty: (1) podkreślenie znaczenia i ważności odgórnego charakteru percepcji; (2) rola kategorii (*hyperpriors*) w poznaniu; (3) ogólna reguła działania modelu generatywnego; (4) proces poznawczy rozumiany jako „analiza przez syntezę"; oraz (5) kluczowa rola wyobraźni w percepcji. Należałoby jeszcze

[26] Zdaniem Kanta John Locke nie odróżnił tych dwóch rodzajów pytania i z tego powodu „[Locke] wykrywając w doświadczeniu czyste pojęcia intelektu, wyprowadzał je także z doświadczenia, a postępował przy tym tak niekonsekwentnie, że odważał się przy ich pomocy na próby wyjaśnienia poznań, które wykraczają daleko poza wszelkie granice doświadczenia" (Kant, 2001, s. 130).

dodać, że zarówno podejście Kanta, jak i to proponowane w ramach PP opierają się na metodzie inżynierii odwrotnej. Jest to jednak problematyczne, co wykazałem wcześniej. Do tego problemu jeszcze powrócę.

Nie wszyscy zwolennicy PP zgadzają się z ogólną konstatacją wskazującą na Kanta jako prekursora i patrona tej koncepcji. Nico Orlandi (2016) twierdzi, że o kantowskim rodowodzie PP mają przemawiać dwa fakty: podkreślenie znaczenia oczekiwań w percepcji oraz stwierdzenie, że percepcja opiera się na inferencjach zstępujących. To drugie oznacza, że percepcja oparta jest na procesie formułowania i testowania hipotez, a jej przedmioty są konstruowane na podstawie wcześniejszych doświadczeń[27]. Takie podejście Orlandi określa nie tyle kantowskim, ile intelektualistycznym i konstruktywistycznym (Orlandi, 2016, s. 12–13, 16). Jej zdaniem można interpretować PP w inny sposób, który określa mianem ekologicznego (por. także Orlandi, 2017)[28].

Jak nietrudno spostrzec, zarówno zwolennicy, jak i przeciwnicy podkreślania kantowskiego charakteru PP, opierają swoje argumenty na tym, co jest wspólne obu podejściom, mając na myśli przede wszystkim literalne podobieństwo pomiędzy pomysłami Kanta a rozwiązaniami przyjętymi w ramach badań nad PP. Ja jednak zaproponuję odmienną strategię. Wielu autorów zwraca uwagę na fakt, że Kant jako pierwszy zaczął stosować tzw. argumenty transcendentalne (por. Hookway, 1999; Walker, 2006). Na czym one polegają? Najogólniej mówiąc, argument transcendentalny polega na wykazaniu, że pewne zakładane przez nas zjawisko zawiera w sobie jako konieczne warunki to, w co albo chcielibyśmy wątpić, albo czego istnienia byśmy *explicite* nie zakładali. W innym sformułowaniu argument transcendentalny polega na wykazaniu, że dane *X* jest koniecznym warunkiem dla jakiegoś *Y* wtedy, gdy można wykazać, że jeżeli *Y* jest faktem, to również *X* jest faktem (Stern, 2019)[29]. Wielu komentatorów, uzasadniając używanie przez Kanta argumentów transcendentalnych, powołuje się na następujący fragment *Krytyki czystego rozumu* (Kant, 2001, s. 580)[30]:

> Dowód nie pokazuje mianowicie, że dane pojęcie (np. tego, co się dzieje) prowadzi wprost do innego pojęcia (np. do pojęcia przyczyny), albowiem takie przejście byłoby skokiem, za który nie można by wcale wziąć odpowiedzialności. Pokazuje on natomiast, że doświadczenie samo, a tym samym przedmiot doświadczenia byłby niemożliwy bez takiego powiązania. Dowód więc musiałby zaznaczyć możliwość dojścia syntetycznie i *a priori* do pewnego poznania rzeczy, które nie byłoby zawarte w ich pojęciach.

---

[27] Wiemy, że nie jest to pomysł nowy, bowiem już Gregory (1980) stwierdził, że percepcja jest w istocie procesem stawiania i uzasadniania hipotez.

[28] Propozycje Orlandi omawiam w podrozdziale *Propozycja Nico Orlandi*.

[29] Prostym przykładem użycia argumentu transcendentalnego jest Austinowskie stwierdzenie: „gdyby wśród rzeczy istniejących nie było niczego innego niż rzeczy zmysłowe, nie bylibyśmy w stanie robić tego, co jesteśmy w stanie robić" (Austin, 1993, s. 51).

[30] Należy zwrócić uwagę na to, że samo pojęcie argumentu transcendentalnego pojawia się w *Krytyce* (Kant, 2001, s. 492).

Uważa się, że argumenty transcendentalne pełnią dwie podstawowe funkcje. Po pierwsze mają służyć do obrony przed sceptycyzmem bądź do jego obalenia. Dokonują tego poprzez wykazanie, że to, co sceptyk podważa bądź w co wątpi, musi istnieć jako warunek konieczny takiego podważania lub wątpienia. Dobrym przykładem takiego rodzaju argumentu jest kartezjańskie *cogito, ergo sum*. Ta antysceptyczna funkcja argumentów transcendentalnych nie jest istotna dla naszych rozważań. Druga funkcja tych argumentów polega na próbie systematycznego badania tego, co dany argument odsłania jako konieczne, a więc pewnych podmiotowych schematów myślenia lub doświadczenia. Można powiedzieć, że funkcja służąca odparciu sceptycyzmu stanowi punkt wyjścia dla zastanowienia się nad tym, co umożliwiło owo odparcie, dlaczego mogło być ono skuteczne (Leszczyński, 2010, s. 51–52). Można by zatem stwierdzić, że głównym celem posługiwania się argumentami transcendentalnymi jest wykazanie prawomocności poznania, czyli wskazania jego warunków możliwości. Argumenty transcendentalne można stosować zarówno w ich funkcji antysceptycznej, jak i poznawczej. Można je również, jak sądzę, stosować wyłącznie w funkcji poznawczej[31]. Tak czynili m.in. Peter F. Strawson oraz John R. Searle.

Powstaje tutaj zasadne pytanie o to, czy można stosować argumentację transcendentalną, rezygnując z transcendentalnej perspektywy? Jest to trudne pytanie i nie będę szukał na nie odpowiedzi w niniejszych rozważaniach. Niemniej niektórzy autorzy sądzą, że jest to dopuszczalne[32]. Jako przykład można podać analizy przeprowadzone przez Quassima Cassama w jego artykule pt. *Self-directed transcendental arguments* (1999). Cassam twierdzi, że argumenty transcendentalne mają dwojaką naturę. Po pierwsze są odkrywcze (*de facto*), ponieważ wykazują warunki możliwości doświadczenia. Po drugie zaś są uprawomocniające (*de iure*), bowiem nie tylko umożliwiają zrozumienie, jak działa poznanie, ale także wykazują, skąd czerpie ono swoją prawomocność. Cassam twierdzi, że obie funkcje argumentów, tj. odkrywcza i uprawomocniająca, nie są wolne od wad i że należy je stosować łącznie. Z tego powodu należałoby uznać, że argumenty transcendentalne nie tyle odkrywają coś czy uprawomocniają, ale po prostu wyjaśniają. Niektórzy uważają, że sam Kant sugeruje takie rozwiązanie, bowiem nie mówi on o dowodzeniu czegoś, co wskazywałoby na funkcję argumentacyjną, ale o omówieniu, co ma wprost sugerować funkcję eksplikacyjną (Leszczyński, 2010, s. 54). Nie jest niestety jasne, co przez wyjaśnianie należy tutaj rozumieć.

Należałoby się teraz zastanowić nad tym, czy o kantowskim charakterze PP nie tyle decyduje podobieństwo pewnych analiz i rozwiązań problemów, co jak wykazuje przekonująco Orlandi, jest ściśle zależne od interpretacji niektórych danych

[31] Taką możliwość sugeruje również Damian Leszczyński.

[32] Wydaje się, że Swanson przyjmuje takie rozwiązanie (2016, s. 2), ale nie czyni z niego argumentu na rzecz kantowskiego charakteru PP. Koncentruje się raczej, co podkreśliłem wcześniej, na literalnych podobieństwach łączących Kanta i PP.

empirycznych, ale to, że w PP mamy do czynienia ze swoistą formą argumentu transcendentalnego. Pomyślmy sobie, że ów argument transcendentalny mógłby mieć następującą postać:

1. Mamy percepcyjne doświadczenie świata, które z natury jest podatne na błędy, złudzenia czy halucynacje.

2. Aby wyjaśnić to doświadczenie, należy wyjaśnić możliwość rozbieżności pomiędzy informacjami dotyczącymi przyczynowych źródeł pobudzeń zmysłowych a faktycznymi przyczynami tych pobudzeń (możliwość minimalizacji błędów predykcyjnych).

3. Jeżeli wyjaśni się możliwość minimalizacji błędów predykcyjnych, to wyjaśni się funkcjonowanie percepcji.

4. Aby wyjaśnić możliwość minimalizacji błędów predykcyjnych, należy przyjąć istnienie hierarchicznego, wielopoziomowego modelu generatywnego.

5. Jeżeli percepcja ma być możliwa, to musi istnieć hierarchiczny, wielopoziomowy model generatywny.

Należałoby się teraz zastanowić, czy takie rozumowanie jest zastosowaniem argumentacji transcendentalnej.

W argumentacji transcendentalnej ma się do czynienia z wykazaniem, że pewne *Y* jest warunkiem koniecznym dla jakiegoś *X*. Taką argumentację Kant wiązał z przeprowadzoną przez siebie dedukcją transcendentalną. Zgodnie z jego intencją dedukcja ta miała być wnioskowaniem, w którym przechodzi się od ogólnych przesłanek do szczegółowego wniosku, opierając się na koniecznym charakterze wiążącym poszczególne zdania. Dedukcja transcendentalna objaśnia sposób, „w jaki pojęcia mogą *a priori* odnosić się do przedmiotów" (Kant, 2001, s. 124)[33]. To, co jest istotne z punktu widzenia przeprowadzonych tu analiz, to fakt, że rozumowanie ma być pewne i konieczne. Oznacza to, że wraz z wykazaniem fałszywości którejś z przesłanek, całe rozumowanie jest fałszywe. Czy mamy do czynienia z takim rodzajem rozumowania w przypadku PP? Trudno jeszcze na to pytanie wprost odpowiedzieć.

Wnioskowania obecne w PP są bayesowskie, co w pełni uzasadnię w dalszej części moich rozważań. Oznacza to, że mają one charakter probabilistyczny. Hipotezy, jakie się tutaj stawia, są więc mniej lub bardziej prawdopodobne, a nie prawdziwe bądź fałszywe. Z tego to też powodu, co podkreśla Anil Seth, bliżej PP do indukcji, która nie ma charakteru koniecznego jak dedukcja, oraz do klasycznego wnioskowania statystycznego (Seth, 2015, s. 2). Bayesowskie wnioskowanie nie jest jednak

[33] Do dziś trwają spory wśród komentatorów, jak właściwie rozumieć kantowską dedukcję transcendentalną. Jak przekonująco wykazał Heinrich (1969), sam Kant miał świadomość, że przeprowadzona przez niego dedukcja mogłaby zostać skonstruowana zupełnie inaczej. Miłkowski podkreśla (2007), że dedukcja transcendentalna jest metodą, która nie określa jedynego możliwego układu warunków: zwykle istnieje wiele różnych przesłanek mogących prowadzić do tego samego wniosku. Każda z nich może dowodzić jego prawdziwości, niemniej żadna z nich nie czyni tego w sposób konieczny.

również indukcyjne, ponieważ nie wnioskuje się tutaj ze szczegółów na temat ogółu, ale raczej ze skutku o jego przyczynie. Hohwy (2014) oraz Seth (2015) stwierdzają, że wnioskowania obecne w PP są abdukcyjne, gdzie abdukcja jest rozumiana jako wnioskowanie do najlepszego wyjaśnienia. Owo najlepsze wyjaśnienie w odniesieniu do bayesowskiego modelowania opiera się na złożoności modelu, uprzedniej wiedzy, tzw. przekonaniach (*prior*) oraz funkcji wiarygodności (*likelihood*).

Abdukcja jest takim wnioskowaniem, w jakim dla pewnego zbioru faktów tworzy się ich najprawdopodobniejsze wyjaśnienie. Struktura tego wnioskowania wygląda następująco: z wniosku *X* oraz przesłanki „jeżeli *Y*, to *X*" wynika przesłanka *Y*. Przykładowo:

Obserwujemy niewyjaśnione zjawisko *X*.
Gdyby *Y* było prawdziwe, to zachodzenie *X* byłoby oczywiste.

Zatem:

Istnieją podstawy, by podejrzewać, że *Y* jest prawdziwe.

Takie rozumowanie można określić mianem „uprawdopodobniającego". Kazimierz Ajdukiewicz wykazał, że do tego typu rozumowań należy także zaliczyć wnioskowania redukcyjne. W *Logice pragmatycznej* (1974) stwierdza, że w tego rodzaju wnioskowaniach przesłanki wynikają z wniosku i z określonego twierdzenia, które dana osoba uważa za prawdziwe. Jest to wnioskowanie odwrotne, ponieważ kierunek, w którym przebiega proces wnioskowania, jest odwrotny do kierunku, w którym przebiega zjawisko. Jest ono również redukcyjne w przeciwieństwie do dedukcyjnego. Polega bowiem na tym, że przesłanki wynikają entymematycznie z wniosków odnoszących się do niektórych stwierdzeń, które są w zakresie wiedzy danej osoby. Twierdzę, że z tego rodzaju wnioskowaniem redukcyjnym mamy do czynienia w PP[34]. Wnioskujemy tutaj bowiem z przesłanek o wniosku, który jest konkluzywny w świetle dotychczas posiadanej wiedzy. Takie wnioskowanie można bezpośrednio wiązać z inżynierią odwrotną. Dla przypomnienia: inżynieria odwrotna to taki sposób postępowania, w jakim analizuje się gotowe urządzenie (przedmiot, strukturę, mechanizm) w celu poznania sposobu jego działania oraz funkcji, które realizuje. Innymi słowy, na podstawie budowy urządzenia wnioskuje się o jego działaniu. Takie rozumowanie ma charakter subiektywny i zawsze istnieje możliwość znalezienia lepszego wyjaśnienia interesującego nas faktu. Zatem wnioskiem wnioskowania redukcyjnego jest również najlepsze wyjaśnienie. Z tego też powodu abdukcję uważam za rodzaj wnioskowania redukcyjnego bądź

[34] W sprawie rozumienia wnioskowań bayesowskich w PP por. rozdział *Przetwarzanie predykcyjne jako bayesowski model eksplanacyjny*, szczególnie podrozdział *Podstawowe pojęcia*.

odwrotnego. Odnosząc się do podanych powodów, należy powiedzieć, że w PP nie wykorzystuje się żadnej argumentacji transcendentalnej. W tym sensie PP nie jest zatem również kantowskie.

Można teraz powrócić do pytania o metodę inżynierii odwrotnej. Ustaliliśmy, że istnieją dwa mocne argumenty odmawiające Kantowi stosowania tej metody. Opierają się na uznaniu, że w metodzie królewieckiego filozofa abstrahuje się od pytań genetycznych oraz od wyjaśnień funkcjonalnych. Pytania *quid juris* należy odróżnić od pytań *quid facti*. Można bronić tezy, że u Kanta istnieje pewna forma transcendentalnie zastosowanej inżynierii odwrotnej. Również w niektórych koncepcjach PP mamy do czynienia z odgórną formą analizy regresywnej, która, jak ustaliliśmy, jest typowa dla inżynierii odwrotnej (por. Griffiths i in., 2010; Hohwy, 2013, s. 8; Swanson, 2016). Nie oznacza to jednak, że PP jest transcendentalne.

Podsumowując ten wątek, należy powiedzieć, że PP nie uzasadnia swoich tez i nie wyjaśnia poznania w sposób transcendentalny[35]. Zatem mówienie o jego kantyzmie, które nie opiera się jedynie na odniesieniu do pewnych „jakościowych" podobieństw pewnych sformułowań używanych przez zwolenników PP z niektórymi twierdzeniami obecnymi w myśli Kanta, jest nieuzasadnione. Fakt, że PP opiera się na zastosowaniu metody inżynierii odwrotnej[36], to jest jednak za mało, by uznać daną dziedzinę za kantowską. Nikt przecież przykładowo nie traktuje pracy programistów jako kantyzmu „w praktyce". Mamy tutaj do czynienia z wnioskowaniem (redukcyjnym) do najlepszego wyjaśnienia. Proponowane ujęcie zrywa zatem z koncepcjami, które utożsamiają „literalnie" PP z projektem autora *Krytyki czystego rozumu* (por. Clark, 2013b; Gładziejewski, 2016; Hohwy, 2013; Swanson, 2016), bądź traktuje je jako kantowskie z powodu jego intelektualizmu (Orlandi, 2016) lub idealizmu (Anderson, Chemero, 2013; Zahavi, 2018)[37]. Samo wykazanie, że PP nie jest kantowskie, nie rozwiązuje jednak problemu dotyczącego różnego rodzaju „metafizycznych" czy „epistemologicznych" konsekwencji PP. Przyjrzę się im bliżej w dalszej części tego rozdziału.

1.4. Krytyka przetwarzania predykcyjnego

Krytyka Zahaviego i innych autorów (por. podrozdział *Możliwe algorytmy przetwarzania predykcyjnego i filozoficzne inspiracje*) wymaga odpowiedzi, ponieważ wydaje się, że przyjęcie ich argumentów może prowadzić do wniosku, iż mamy do

[35] Teza ta znajduje swoje dodatkowe uzasadnienie w rozdziale *Przetwarzanie predykcyjne jako bayesowski model eksplanacyjny*, w którym omawiam rozumienie podstawowych pojęć bayesowskiego PP.

[36] W rozdziale *Przetwarzanie predykcyjne jako bayesowski model eksplanacyjny* analizuję, z jakiego rodzaju inżynierią odwrotną ma się do czynienia w PP, oraz co to w praktyce badawczej oznacza.

[37] Można bronić kantyzmu PP jedynie w tym szerokim sensie, w jakim za kantowskie uważa się wszelkie teorie czy modele, które uznają różne kategorie, pojęcia czy struktury określające charakter percepcji i poznania za wrodzone.

czynienia nie tylko z koncepcją odnawiającą stare spory epistemologiczne, ale także może prowadzić do przeświadczenia, że jesteśmy skazani na wiele metafizycznych aporii sięgających XIX wieku. Przyjrzymy się teraz bliżej zarzutowi Zahaviego[38]. Zgodnie z jego ujęciem PP prowadzi do radykalnego neuroreprezentacjonizmu i internalizmu. Zahavi twierdzi, że według PP „treść naszych świadomych doświadczeń jest konstrukcją neuronalną, symulacją generowaną przez mózg", co oznacza, że „nasz dostęp do świata zewnętrznego jest zapośredniczony przez reprezentacje neuronalne, zaś świat doświadczenia jest reprezentacyjnym konstruktem" (Zahavi, 2018, s. 47–48). Według Zahaviego PP nie uzasadnia bezpośredniego kontaktu percepcyjnego ze światem zewnętrznym, ponieważ jak twierdzi „bezpośrednimi przedmiotami percepcji są w rzeczywistości konstrukcje mentalne. Spostrzegana róża, dotknięta kostka lodu, usłyszana melodia itd. są wewnętrznymi reprezentacjami generowanymi przez mózg". W związku z tym Zahavi dopytuje: „Dlaczego raczej rozważamy możliwość, że przedmioty naszej percepcji są konstruktami generowanymi wewnętrznie, a nie prawdziwymi obiektami czasoprzestrzennymi? Robimy tak, ponieważ tak sugerują nasze neuronaukowe badania mózgu" (Zahavi, 2018, s. 49, 54). Oznacza to, że internalizm i reprezentacjonizm obecny w PP mają wprost prowadzić do sceptycyzmu i postawy antyrealistycznej.

Podsumowując krytykę Zahaviego, należy stwierdzić, że PP nie uwzględnia „naszej naturalnej skłonności do realizmu", ponieważ nie zakłada realizmu naukowego. Zahavi zastanawia się: „Skąd możemy wiedzieć, że nasze teorie naukowe, które są oferowane jako wyjaśnienia świata doświadczeń, naprawdę docierają do rzeczywistości zewnętrznej?"; „W jaki sposób naukowcy mogą przekroczyć symulację swojego wewnętrznego świata? Jak udaje im się uzyskać to epistemiczne osiągnięcie?"; „W jaki sposób omawiane stanowisko (= PP) może przyjąć realizm naukowy?" (Zahavi, 2018, s. 53).

Twierdzę, że cała krytyka Zahaviego sprowadza się do kluczowego pytania o to, w jaki sposób rama PP może obronić „naszą naturalną skłonność do realizmu" (Zahavi, 2018, s. 47), którą ten autor wiąże ze stanowiskiem naiwnego realizmu, czyli „naszego przekonania o obiektywnym istnieniu przedmiotów codziennego doświadczenia" (Zahavi, 2018, s. 54).

W niniejszym rozdziale będę twierdził, jednocześnie odrzucając pogląd Zahaviego, że PP nie zakłada ani nie prowadzi do konsekwencji oraz rozwiązań antyrealistycznych. Rozumiem przez to, że obecne w PP założenia reprezentacjonistyczne oraz inferencjalizm nie przesądzają sprawy na rzecz tezy, iż nie mamy bezpośredniego dostępu do rzeczywistości zewnętrznej, a przedmioty takie jak mózg czy świat są jedynie konstruktami mentalnymi. Uważam raczej, że istnieją mocne przesłanki, aby twierdzić, że reprezentacjonizm i inferencjalizm PP nie tylko nie wspierają tezy antyrealistycznej, ale przeciwnie: pozwalają one na uzasadnienie tezy realistycznej,

[38] Wstępną krytykę ujęcia Zahaviego przeprowadziłem w: Piekarski, 2017b.

zgodnie z którą istnieje zewnętrzny i obiektywny świat wyjaśniany przez nasze teorie naukowe. W tym sensie będę twierdził, że PP jest zgodne ze stanowiskiem realizmu. Rozpocznę moje omówienie od wykazania, że krytyka PP jako swoistej formy idealizmu, konstruktywizmu albo neuroreprezentacjonizmu opiera się na bardzo selektywnej lekturze różnych interpretacji PP. Mianowicie: krytykowani są tylko ci badacze, których można zaliczyć do tzw. konserwatywnego PP. Dla tego ujęcia istotny jest reprezentacjonizm i internalizm[39].

Następnie sformułuję odpowiedź na antyrealistyczny zarzut Zahaviego, omawiając znaczenie internalizmu w PP oraz rolę, jaką grają w PP realne wzorce (por. Dennett, 2008) oraz reprezentacje strukturalne (S-reprezentacje) (por. Gładziejewski, 2016; Gładziejewski, Miłkowski, 2017).

1.4.1. Konserwatywne przetwarzanie predykcyjne

Już pobieżna lektura niektórych opracowań dotyczących PP prowadzi do konkluzji, że nie ma jednej i wyczerpującej wykładni tej koncepcji. Przeciwnie istnieje wiele różnych, często wykluczających się i krytykujących, ujęć. Analiza tej koncepcji pozwala dostrzec szereg rozbieżności dotyczących rozumienia podstawowych pojęć, terminologii, statusu eksplanacyjnego czy metodologicznego poszczególnych twierdzeń, a także roli, jaką może ona odgrywać w korpusie nauk o poznaniu i filozofii. Wyróżnia się dwa podstawowe podejścia: konserwatywne i radykalne (por. Clark, 2015a, 2015b; Gładziejewski, 2017b; Orlandi, Lee, 2018). Odróżnienie tych ujęć jest nie tylko ważne z punktu widzenia celów przewidzianych do zrealizowania w tej pracy, ale także dla wykazania niezasadności wielu krytyk PP.

W konserwatywnym podejściu do PP (por. Gładziejewski, 2016, 2017b, 2018; Hohwy, 2013; Kiefer, Hohwy, 2017; Wiese, 2016a) podkreśla się, że umysł jest względnie izolowany od środowiska, co oznacza, że nasz poznawczy kontakt ze światem zewnętrznym ma swoje źródło w neuronalnej aktywności mózgu i tylko w niej. Model świata jest kodowany na poziomie neuronalnym (Clark, 2015a, s. 14). Model generatywny wytwarza predykcje na podstawie wewnętrznych reprezentacji (S-reprezentacji) środowiska zewnętrznego (Gładziejewski, 2016). Znaczy to, że relacja pomiędzy umysłem a światem jest zapośredniczona przez wewnętrzne reprezentacje. Stanowisko takie można za Andersonem określić jako „rekonstrukcyjne" (Anderson, 2014, ponieważ w tym ujęciu model świata jest rekonstruowany na podstawie wewnętrznych reprezentacji świata, które są izomorficzne z przyczynową strukturą świata (Gładziejewski, 2016; Kiefer, Hohwy, 2018). Niektórzy określają te reprezentacje jako strukturalne, ponieważ działają

[39] Trzeba jednak podkreślić, że przedstawiona przez niego krytyka przede wszystkim dotyczy konserwatywnego PP, zaś w znacznie mniejszym stopniu jego antyreprezentacjonistycznej alternatywy (Zahavi, 2018, s. 55).

one na podstawie podobieństwa strukturalnego zachodzącego pomiędzy samą reprezentacją a tym, co reprezentowane (por. Gładziejewski, 2016; Gładziejewski, Miłkowski, 2017; Kiefer, Hohwy, 2018; O'Brien, Opie, 2004; Shea, 2014). W tym kontekście można przywołać słowa Hohwy'ego, który twierdzi, że „przyczynowa struktura środowiska i przyczynowa struktura reprezentowana w wewnętrznym modelu, będą się wzajemnie odzwierciedlać" (Hohwy, 2018, s. 4), ponieważ system poznawczy jest „wewnętrznym lustrem natury" (Hohwy, 2013, s. 220).

Paweł Gładziejewski (2017b) wskazuje na trzy podstawowe zobowiązania charakterystyczne dla konserwatywnego podejścia do PP. Są to:

1. Zobowiązanie do reprezentacjonizmu.
2. Zobowiązanie do używania pojęcia inferencji jako wiążącej percepcję i działania.
3. Zobowiązanie do internalizmu jako stanowiska zakładającego, że mechanizmy poznawcze oparte są wyłącznie na pracy centralnego układu nerwowego. Oznacza to, że treść reprezentacji mentalnych jest wyznaczana jedynie przez stany wewnętrzne danego organizmu (Lau, Deutsch, 2002).

Dla analiz prezentowanych w tej pracy szczególnie istotne jest pierwsze zobowiązanie. Gładziejewski twierdzi, że można je interpretować na dwa sposoby. Tak zwana słaba interpretacja zakłada, że treść stanów mentalnych jest przypisywana wewnętrznemu modelowi świata z czysto pragmatycznych powodów. W tym ujęciu wewnętrzne reprezentacje nie mają realnej treści, ale są jedynie postulowane przez niektórych badaczy, którzy chcą wyjaśnić poznawcze funkcje umysłu (por. Downey, 2017; Egan, 2014)[40]. W mocnym ujęciu konserwatywnego podejścia do PP postuluje się, że stany mentalne zawierają realną oraz skuteczną przyczynowo treść reprezentacyjną. Z tego powodu należy stwierdzić, że model generatywny ma faktyczną treść i w sposób nietrywialny reprezentuje świat. Oznacza to, że (1) generuje on takie przewidywania dotyczące środowiska, jakie kierują działaniami; (2) funkcja kierowania działaniami jest zależna od podobieństwa funkcjonalnych relacji zachodzących pomiędzy reprezentacjami a przyczynową strukturą środowiska; (3) stopień strukturalnego podobieństwa pomiędzy środowiskiem a modelem jest zależny przyczynowo od stopnia skuteczności modelu w działaniu.

Przyjmując mocną interpretację, należy stwierdzić, że zobowiązanie do inferencyjnego charakteru relacji pomiędzy percepcją a działaniami oraz zobowiązanie do inferencjalizmu są ściśle zależne i wynikają ze zobowiązania do reprezentacjonizmu (Gładziejewski, 2017b, s. 106, 111; Kiefer, 2017)[41]. Krytycy tego ujęcia (Bruineberg, Kiverstein, Rietveld, 2016; Orlandi, 2016, 2017; Zahavi, 2018) twierdzą, że internalizm oznacza takie stanowisko, jakie w przeciwieństwie do rozszerzonych

[40] Krytykę tego rodzaju „fikcjonalistycznych" podejść przeprowadził Mark Sprevak (2013).

[41] Chodzi o to, że mózg abdukcyjnie „wnioskuje" o środowiskowych przyczynach pobudzeń zmysłowych (*sensory input*) w taki sposób, że przedstawia hipotezy, które najlepiej wyjaśniają informacje pochodzące ze środowiska (Kiefer, 2017; zob. także Gregory, 1980).

i ucieleśnionych podejść do poznania podkreśla, iż granicą, w obrębie której należy myśleć i badać poznanie, jest granica układu nerwowego. W ten sposób mocna interpretacja konserwatywnego ujęcia PP miałaby zakładać przesłanki indywidualizmu metodologicznego, tj. takiego stanowiska, jakie zakłada, że wyjaśnienie procesów i zjawisk poznawczych odbywa się poprzez analizę i wyjaśnienie mechanizmów poznawczych realizowanych w mózgu, w konsekwencji czego w znacznym stopniu ignoruje się środowisko oraz interakcje z nim, gdyż nie są to własności indywiduum (por. Fiske, Taylor, 2008). W przeciwieństwie do takiej interpretacji będę twierdził, że mózg, a w szczególności model generatywny, rekapituluje statystyczną strukturę środowiska. Oznacza to, że środowisko odgrywa pośrednią, ale absolutnie niezbędną, rolę w wyjaśnieniach psychologicznych oferowanych przez tzw. konserwatywne PP. Teza ta zostanie uzasadniona w dalszych analizach.

1.4.1.1. Internalizm a koce Markowa

Można teraz zrozumieć, dlaczego Zahavi twierdzi, że konserwatywne PP prowadzi do antyrealizmu i „kartezjańskiego internalizmu" (Zahavi, 2018, s. 57). Skoro bowiem, jak sugerują niektórzy autorzy, PP jest reprezentacjonistyczne, zaś model generatywny jest internalistyczny, czyli jego granicą jest granica układu nerwowego, to zasadny jest sceptycyzm w stosunku do przekonania głoszącego realność świata zewnętrznego, bowiem kontakt poznawczy ze światem ma swoje źródło tylko w mózgu i jest zapośredniczony przez neuroreprezentacje. Twierdzę jednak, że argumentacja Zahaviego i zwolenników radykalnego PP nie jest przekonująca.

Należy zapytać o to, co czyni model generatywny internalistycznym? Hohwy podkreśla, że myślenie o systemach wewnętrznych (*internal systems*) wywodzi się z uczenia maszynowego oraz nauk obliczeniowych i jako takie nie może być w sposób bezpośredni zastosowane do opisu organów biologicznych. Jego zdaniem granice systemów żywych najlepiej wyznaczają ich koce Markowa[42] (Hohwy, 2017; por. także Hesp i in., 2019; Kirchhoff i in., 2018). W tym ujęciu należy powiedzieć, że stany wewnętrzne danego organizmu można zdefiniować jako koce Markowa, które oddzielają system od jego środowiska. Stany te są zatem niezależne (w sensie statystycznym) od stanów otoczenia. Oznacza to, że stany wewnętrzne mają za zadanie minimalizować błędy predykcyjne, aby zachować strukturalną i funkcjonalną integralność, która prowadzi do homeostazy i samoorganizacji. Warunkiem

[42] Za pomocą koców Markowa (Pearl, 1988, s. 120–121) można opisać taką serię węzłów, że dla danego węzła *A* jego zachowanie można w pełni przewidzieć, znając stany innych węzłów, które tworzą właściwy mu koc. Stany sąsiadujących węzłów określają w sposób statystyczny (nie przyczynowy) stan docelowego węzła *A*. Znaczy to, że koc Markowa dla węzła *A* zawiera całą wiedzę potrzebną do przewidywania zachowania tego węzła i węzłów wynikających zeń. Formalne pojęcie koców Markowa i powiązane z nim pojęcie sieci bayesowskiej wprowadzam w podrozdziale *Podstawowe pojęcia*. Dodatkowe ich omówienie por. podrozdział *Koce Markowa a przetwarzanie predykcyjne*.

minimalizacji jest zatem ścisła separacja wnętrza systemu od jego otoczenia, czyli zdolność tego systemu do odróżnienia siebie od środowiska (Friston, 2013a). Koce Markowa umożliwiają statystyczne wyznaczenie granic systemu. Pojawia się podział pomiędzy stanami wewnętrznymi (systemowymi) i zewnętrznymi (środowiskowymi). Te ostatnie mogą być wiązane ze stanami neuronalnymi, stany ciała bądź otoczenia, w zależności od podziału danego systemu. Koc obejmuje podział na stany aktywne i sensoryczne, które pośredniczą w wymianach pomiędzy stanami systemowymi i środowiskowymi (neuronalnymi, cielesnymi, światowymi). Hohwy twierdzi, że istnieje zasadnicza różnica pomiędzy znanymi przyczynami wewnętrznymi, ponieważ są one wywnioskowane przez model, a zewnętrznymi, ukrytymi przyczynami znajdującymi się po drugiej stronie koca Markowa. Zgodnie z tym ujęciem internalizm oznacza, że całe przetwarzanie percepcyjne i poznawcze odbywa się w modelu wewnętrznym lub, innymi słowy, w obrębie koca Markowa (Hohwy, 2017, s. 7). Należy zatem powiedzieć, wbrew tezie bronionej przez Zahaviego, że w konserwatywnym PP kontakt poznawczy zależy od tego, gdzie jest umieszczony koc Markova, którego lokalizacja jest niezależna od jakichkolwiek faktów na temat mózgu, wzorców pobudzeń neuronalnych itd. To, że np. Hohwy twierdzi, iż koc najlepiej umieścić wokół mózgu (Hohwy, 2014, s. 18), jest ustaleniem warunkowym podyktowanym m.in. strategią eksplanacyjną, zainteresowaniami badawczymi, jak i samym badanym zjawiskiem[43]. Z tego też powodu odrzucam tutaj pogląd określony przez Ramsteada i in. autorów (2019) jako esencjalizm, czyli przekonanie, zgodnie z którym „istnieje unikatowa granica lub jednostka analizy, z której najlepiej zrozumieć i zbadać poznanie" (Ramstead i in., 2019, s. 2)[44].

Powyższe analizy prowadzą do stwierdzenia, że można być w PP internalistą, jednocześnie nie będąc neuroreprezentacjonistą. Wszystko zależy od tego, gdzie umieści się koc Markowa. Innymi słowy, internalizm nie musi implikować i faktycznie nie implikuje neuroreprezentacjonizmu. Z tego powodu można zgodzić się z Gładziejewskim, który twierdzi, że konsekwentne odczytanie podejścia konserwatywnego (nawet przy mocnej interpretacji zobowiązania do reprezentacji) pozwala włączyć je do nurtu 4E nad poznaniem. Za tym twierdzeniem przytacza wiele argumentów. Po pierwsze reprezentacje nie są statycznymi obrazami rzeczywistości, ale wewnętrznymi, przewodzącymi działanie, reprezentacjami strukturalnymi, które pozwalają na rozpoznanie błędów reprezentacyjnych. Mają

[43] Warto dodać, że z perspektywy wyjaśniania mechanistycznego (por. podrozdział *Mechanistyczny model wyjaśniania naukowego a mechanizmy normatywne*) wyznaczenie granic danego systemu poznawczego zawsze jest uzależnione od danego zadania badawczego – a więc tego, w jaki sposób dane zjawisko poznawcze chcemy wyjaśnić, tym samym wskazując mechanizm za nie odpowiedzialny. Wyznaczenie granic systemu poznawczego odbywa się więc za każdym razem od nowa, bez żadnych przesądzeń co do jego szerokości czy scentralizowania wokół mózgu, za to stosując mechanistyczne kryterium odróżniania komponentów systemu od warunków kontekstowych (Wachowski, 2018b; por. także Hutchins, 2014; Kaplan, 2012).

[44] Odrzucenie esencjalizmu w kwestii granic systemów poznawczych może prowadzić do osłabienia bądź nawet zanegowania ostrego rozróżnienia pomiędzy internalizmem i eksternalizmem (por. na ten temat Ramstead i in., 2019).

one charakter modalny, a ich treść jest ograniczana (*constrained*) przez sposób, w jaki organizm jest ucieleśniony i osadzony w środowisku. Po drugie kluczowe dla PP pojęcie inferencji oznacza, że reprezentacje mają warunki prawdziwości, a sposób, w jaki są one aktualizowane, jest aktywny, a nie reaktywny. Po trzecie zaś zasoby pojęciowe związane z PP pozwalają na interesujące powiązanie tego ujęcia z innymi istniejącymi koncepcjami z zakresu 4E. Hohwy (2018) dodaje, że PP pozwala lepiej wytłumaczyć sposób, w jaki ucieleśnione podmioty wchodzą w dynamiczne interakcje ze środowiskiem, oraz że granica pomiędzy umysłem i światem z jednej strony ma charakter samouwierzytelniający się (*self-evidencing*) (Hohwy, 2014), co oznacza, że przyczyny pobudzeń zmysłowych znane są pośrednio przez wnioskowanie na temat informacji docierającej z wejść zmysłowych, z drugiej zaś ma ona charakter przyczynowy. Istnieje dynamiczne sprzężenie zwrotne pomiędzy umysłem a światem, możliwe dzięki percepcji oraz działaniom realizowanym w tym świecie.

1.4.1.2. Problem sceptyczny

Należy stwierdzić, że opis modelu w kategoriach koca Markowa i jego węzłów pozwala wyjaśnić samouwierzytelnianie się modelu. Zgodnie z PP model generatywny, którego granice określa jego koc Markowa, wnioskuje (w sensie rekapitulowania przyczynowej struktury świata)[45] o ukrytych przyczynach sygnału zmysłowego odbieranych przez wejścia zmysłowe. Jest to kluczowe dla modelu, ponieważ w ten sposób maksymalizuje swoją wierzytelność (*self-evidence*), którą można rozumieć jako oczekiwaną nagrodę dla przyszłych stanów podmiotu (por. Schwartenbeck i in., 2013 oraz podrozdział *Odpowiedzi zwolenników przetwarzania predykcyjnego i zasady energii swobodnej*). Możemy dodać, że węzły wewnętrznego modelu generatywnego mają dostęp do stanów koca i są w stanie zminimalizować potencjalne błędy predykcyjne (Hohwy, 2017, s. 4). W ten sposób model minimalizujący potencjalne błędy predykcyjne oczekuje, że informacje docierające do mózgu będą do pewnego stopnia wiarygodne (*reliable*). Gdyby tak nie było, aktualizacja modelu świata nie byłaby możliwa. Chodzi zatem o to, aby zapobiec nadmiernemu zniekształceniu modelu przez niewiarygodne dane. Oczywiście stany świata są ukryte przed obserwatorem, więc to, co bezpośrednio tłumaczy się we wnioskowaniu percepcyjnym, to wzorce statystyczne rozpoznawalne na stymulacji wejść zmysłowych. Wyjaśnienie stale zmieniających się, ale w miarę łatwo przewidywalnych wzorców danych zmysłowych umożliwia wewnętrznemu modelowi wybieranie i kształtowanie własnych parametrów. Owo zwiększanie własnej wierzytelności

[45] Nie jest to zatem wnioskowanie, o czym nieustannie należy tutaj pamiętać, w sensie potocznym jako świadome przechodzenie od przesłanek (określonych zdań) do wniosków (innych zdań).

odbywa się w ścisłym związku przyczynowym ze światem: ukryte przyczyny pobudzeń zmysłowych muszą zostać wywnioskowane we wnioskowaniu percepcyjnym w celu skutecznego działania podmiotu w środowisku. Jest to możliwe dzięki temu, że odpowiednie ukryte zmienne w modelu kodują rozkłady prawdopodobieństwa oparte na obserwacji środowiska i w odniesieniu do niższych poziomów modelu. W tym sensie można stwierdzić, że przyczyny kształtujące treść i charakter sygnału zmysłowego są reprezentowane w kategoriach prawdopodobieństwa wytworzenia określonych wzorców sensorycznych w systemie. Określonym relacjom pomiędzy zmiennymi (parametrami) modelu odpowiadają w pewnym stopniu określone relacje diachroniczne w środowisku (por. Gładziejewski, 2016, s. 571–572). To, że np. samochód jest percypowany w taki, a nie inny sposób, wiąże się z tym, że niższy poziom modelu odpowiedzialny za detekcje krawędzi odnosi się do zmian luminacji w otoczeniu, czyniących określone kontury przedmiotu mniej lub bardziej wyraźnymi, zaś inny, wyższy poziom modelu realizujący procesy rozpoznawania kształtów wiąże się z określonymi własnościami przedmiotu umieszczonego w przestrzeni albo na jakiejś płaszczyźnie. Z tego też powodu zmiany kształtu, obroty i zmiany skali danego przedmiotu odpowiadają bezpośrednim przekształceniom określonych zmiennych bądź ich konfiguracji w modelu (por. Ballard, 1981).

Hohwy podkreśla, że zakładamy, iż wzorcom danych sensorycznych odpowiadają faktyczne stany świata oraz ciała, niemniej, co uwyraźnia krytyka Zahaviego, sytuacja ta może rodzić wątpliwości sceptyczne. Ktoś mógłby np. twierdzić, że przyczyną informacji zmysłowej nie są znane nam stany świata, ale złośliwy demon lub złośliwy naukowiec (Hohwy, 2013, s. 179, 2017, s. 4). Gdyby taki demon istniał, to nie moglibyśmy się dowiedzieć o tym, ponieważ jego działanie polegałoby na ciągłym wprowadzaniu nas w błąd. Faktyczne przyczyny pobudzeń zmysłowych nie byłyby odróżnialne od tych generowanych przez demona[46]. Zgodnie ze sformułowanym przez Zahaviego zarzutem antyrealizmu PP nie może zagwarantować, że cokolwiek w prawdziwym świecie zewnętrznym odpowiada parametrom wewnętrznego modelu. Ujęcie to jest zgodne z zarzutem sceptyka.

Ustaliliśmy już, że charakter i treść parametrów modelu są wytwarzane w ścisłym związku przyczynowym ze światem, co w przypadku PP oznacza, iż ukryte przyczyny wzorców danych sensorycznych muszą zostać ujawnione we wnioskowaniu przyczynowym. Podejmowane przez agenta działania są zależne od owego wnioskowania. Działanie, zgodnie z ujęciem Fristona (2010), na które powołuje się Hohwy, polega na selektywnym próbkowaniu świata w ramach aktywnego wnioskowania[47]. Gdyby złośliwy demon istniał, to miałby trudne do wykonania

[46] Hohwy podkreśla, że na sceptycyzm podatne jest zarówno konserwatywne, jak i radykalne PP (2017). Tego poglądu nie podziela Clark (2017).

[47] Jest to takie działanie, jakie ingeruje w przyczynową strukturę świata w celu zminimalizowania potencjalnego błędu predykcyjnego (Friston, 2010, s. 129). Por. na ten temat rozdział *Mechanizmy normatywne działania w przetwarzaniu predykcyjnym*.

zadanie, które polegałoby na ciągłym śledzeniu tego selektywnego próbkowania. Musiałby wciąż wytwarzać takie przyczyny, jakie odpowiadałyby faktycznym działaniom podmiotów. Z tego powodu powinniśmy twierdzić, że to, w jaki sposób odzwierciedlamy przyczynową strukturę świata, jest ściśle zależne od procesów wnioskowania, które są realizowane w mózgu (Hohwy, 2013, s. 220–221). Należy jednak zauważyć, co podkreśla Hohwy, że skuteczność procesu minimalizacji błędu predykcyjnego zależna jest tylko od informacji o przyczynowo statystycznych własnościach danych wejściowych, a nie od wewnętrznych aspektów samych przedmiotów. Tylko działanie w świecie (aktywne wnioskowanie) w pewnym stopniu gwarantuje, że mamy możliwość nie tylko zmieniać parametry własnego modelu, ale również ingerować w strukturę przyczynową świata. Wciąż jednak sceptyk może twierdzić, że ukryte przyczyny informacji zmysłowej ostatecznie mogą być ukrytymi pragnieniami demona (Hohwy, 2017, s. 12). Jednak zdaniem Hohwy'ego znacznie lepiej jest żyć z ryzykiem radykalnego sceptycyzmu, niż twierdzić, że nasza percepcja i działania są jedynie skutkami naszych wniosków dotyczących statystycznych własności wkładu sensorycznego i że nie mają żadnych określonych mocy przyczynowych. Twierdzę zatem, że hipoteza Hohwy'ego daje lepsze wyjaśnienie zagadnienia percepcji niż hipoteza sceptyczna, czyli mówiąc innymi słowami, „hipoteza świata rzeczywistego" (tzn. że istnieją stoły, komputery, inni ludzie itd., tj. przedmioty, które *powodują* nasze doświadczenia) daje lepsze wyjaśnienie naszych doświadczeń niż hipoteza złośliwego demona (por. Vogel, 1990).

Odpowiedź Hohwy'ego może nie zadowolić sceptyka. Niemniej można się zastanawiać nad tym, czy łagodzi ona zarzut Zahaviego, jakoby świat doświadczania był tylko reprezentacyjnym konstruktem (Zahavi, 2018, s. 48). Skoro nie wiemy, jakie są przyczyny pobudzeń zmysłowych, to być może takich przyczyn w ogóle nie ma, a świat doświadczenia jest ułudą? Twierdzę, że Zahavi nie ma słuszności. Aby to wykazać, dokonam analizy roli rozpoznawania wzorców w PP.

1.4.1.3. Rzeczywiste wzorce w przetwarzaniu predykcyjnym

Współczesne badania z zakresu kognitywistyki oraz uczenia maszynowego (por. Bishop, 2006) pozwalają stwierdzić, że procesy percepcyjne, uczenia się, nabywania języka, rozpoznawania twarzy itd. opierają się na zdolności do rozpoznawania wzorców (*pattern recognition*). Skrótowo rzecz ujmując, można stwierdzić, że percepcja jest możliwa dzięki temu, iż istnieją pewne stałe, regularne i nielosowe elementy w doświadczeniu zmysłowym (Dennett, 2008), które pozwalają organizować nam dane obecne w statystycznym sygnale zmysłowym. Dzięki wzorcom informacje docierające do mózgu za pośrednictwem modalności zmysłowych są w jakimś stopniu uporządkowane i – co szczególnie istotne – można na ich podstawie prze-

widywać dalszy przebieg danego doświadczenia. Należy zatem stwierdzić, że do pełnego wyjaśnienia procesów predykcyjnych konieczna jest analiza zdolności do rozpoznawania wzorców oraz natury samych wzorców. Z całą pewnością pomiędzy umiejętnością tworzenia hipotez percepcyjnych a zdolnością do rozpoznawania wzorców istnieje pewnego rodzaju relacja zwrotna, ale dotychczasowe analizy prawie nic na temat natury tej relacji nie mówią. Istotne jest, co podkreśla Hohwy, że na wyższych poziomach modelu generatywnego system poznawczy spostrzega przede wszystkim wzorce, a nie indywidualne różnice (Hohwy, 2013, s. 69).

Problem rozpoznawania i natury wzorców jest kluczowy dla przeprowadzanych przeze mnie analiz, ponieważ (1) – co już podkreślałem – podmiot dobiera działania, uwzględniając określone parametry modelu w odniesieniu do wzorców rozpoznawalnych w sygnale zmysłowym, oraz – za czym będę argumentował – (2) pozwala na uzasadnienie pewnej formy realizmu w odniesieniu do PP. Moje analizy rozpocznę od wyjaśniania mechanizmu rozpoznawania wzorców przez model generatywny i jego znaczenia dla działań podejmowanych przez podmiot w środowisku.

Przez wzorzec, najogólniej ujmując, można rozumieć taką własność zbioru danych, jaką można wykryć lub rozpoznać w tych danych (Dennett, 2008; Ströing, 2018, s. 102)[48]. Tak rozumiany wzorzec może być dostępny epistemicznie w formie fizycznej bądź matematycznej reprezentacji. W takim sensie dane empiryczne można opisać w terminach rozpoznawalnych wzorców i odróżnionego od nich szumu. Dobrze to przedstawia rycina zaproponowana przez Daniela Dennetta (2008, s. 303).

Rysunek 2. Graficzne przedstawienie ramek zawierających skończone zbiory danych z określoną (procentowo) ilością szumu (Dennett, 2008, s. 303)

[48] Dennett (2008) stwierdza, że wzorzec jest „z definicji" tym, co może być rozpoznane jako wzorzec. Jak go jednak wykryć? Autor ten odrzuca takie pytania jako źle postawione, bowiem o wzorcowości nie decydują jakieś faktyczne przedmioty czy ich aspekty dostrzegalne w środowisku, ile istnienie opisu, który będzie wydajniejszy od szczegółowego zdania relacji z czasowego przepływu danych w sygnale zmysłowym. Jeżeli taki opis jest możliwy, to mamy do czynienia z rozpoznaniem odpowiedniego wzorca, który możemy wprost związać z rozpoznawaniem pewnych regularności. Dennett zatem proponuje, aby traktować opis jako swoistą kompresję danych. Kompresji może jednak podlegać to, co nielosowe, czyli regularne i powtarzalne. To, co nielosowe jest kompresowalne w pewną skończoną formułę, w konkretny algorytm. Zaś tego, co losowe, skompresować się nie da. Realne wzorce są dla Dennetta kompresowalnymi danymi. Rozpoznawanie wzorców jest rozpoznawaniem pewnych regularności, które można ująć w formie jakiegoś skończonego opisu, w formie algorytmicznej.

Ramki przedstawiają ten sam wzór z różnym natężeniem szumu. Współczynnik szumu przedstawia się następująco:

A: 25% B: 10%
C: 25% D: 1%
E: 33% F: 50%

Kod paskowy jest konkretyzowany w każdej z ramek jako specyficzny wzorzec, który jest łatwo rozpoznawalny dla ludzkiego aparatu wzrokowego, dzięki konstytutywnym dla naszej możliwości widzenia detektorom jaskrawości, głębi czy krawędzi. Dennett stwierdza, że wobec tak rozumianych wzorców możemy przyjąć trzy różne postawy: (1) Czasem chodzi nam o dokładny opis danego wzorca bez względu na koszty. W tym przypadku wzorzec w danej ramce to taki, a nie inny kod paskowy o określonych parametrach. (2) Czasem chcemy opisać szum oraz jego umiejscowienie. W tym ujęciu wzór w ramce A to kod paskowy z 25% szumu. (3) Innym razem ignorujemy poziom szumu i wtedy możemy stwierdzić, że wzorcem w ramce A jest kod paskowy (Dennett, 2008, s. 307–308). Zdaniem Dennetta możliwość ujmowania wzorców na każdym z przynajmniej tych trzech sposobów przemawia na rzecz ich realności. Oznacza to, że wzorce istnieją niezależnie od tego, czy są dla nas dostrzegalne, czy nie, tzn. organizmy, uzbrojone w inne organy zmysłowe niż my, mogłyby je bez trudu rozpoznać. Teraz należy przyjrzeć się bliżej samemu problemowi rozpoznawalności wzorców i jego możliwości w odniesieniu do PP.

Na początek należy powiedzieć, że sygnał zmysłowy można opisać w terminach rozpoznawalnych wzorców i szumu. Takie ujęcie sugeruje dwie interpretacje: albo (1) tylko niektóre wzorce, które możemy rozpoznać w określonych zbiorach danych (strukturze sygnału zmysłowego), odpowiadają faktycznym strukturom w świecie, następnie zaś oszacować, w jakim stopniu wzorce odpowiadające strukturom w świecie różnią się od wzorców, które im nie odpowiadają; lub (2) stwierdzić, że wszystkie wzorce rozpoznawalne w sygnale zmysłowym odpowiadają strukturom obecnym w świecie.

Istotnym ograniczeniem nie tylko dla PP, co już wiemy, ale również, jak podkreśla wielu badaczy (por. McAllister, 2007, 2009; Woodward, 2009), nauk empirycznych jest fakt, że nie możemy odwoływać się do bezpośredniej wiedzy o tym, jakie struktury zawiera świat. Gdybyśmy mogli tak po prostu stwierdzić, że wzorce rozpoznawalne w sygnale zmysłowym mają znaczenie dowodowe, to twierdzilibyśmy, że odpowiadają one określonym i weryfikowalnym strukturom w świecie. Niestety, o tych strukturach możemy tylko wnioskować na podstawie dostępnych wzorców. Musimy zatem wykorzystać te informacje, które są nam dostępne, ponieważ dane w sygnale zmysłowym stanowią nasz jedyny epistemiczny

dostęp do świata i jego struktur. Należy zatem odróżnić wzorce, które odpowiadają określonym strukturom w świecie, od tych, które nie są zgodne z danymi empirycznymi (McAllister, 2009, s. 76). Te pierwsze wzorce mają dla nas siłę dowodową, ponieważ zwiększają wierzytelność modelu, a zwiększają wierzytelność modelu, ponieważ, jak twierdzę, odpowiadają im jakieś struktury w świecie. Wzorce, które nie mają siły dowodowej, są zatem nieistotne ze względu na ilość szumu itd. Z tego wynika, że musimy być w stanie określić różnicę pomiędzy wzorami, które odpowiadają strukturom na świecie, a wzorami, które nie mają znaczenia dowodowego, czyli nie są realne. Jeżeli Zahavi ma rację w swojej krytyce konserwatywnego PP, to model generatywny albo (1) nie jest w stanie odróżnić wzorców mających siłę dowodową od tych, które tej siły są pozbawione; albo (2) nie może uzasadnić, że jakiekolwiek wzorce mają siłę dowodową.

Należy teraz doprecyzować, co rozumiem przez siłę dowodową wzorca. Wzorzec ma siłę dowodową lub innymi słowy, odpowiada określonym strukturom w świecie, jeżeli jest przyczynowo skuteczny oraz ma typowe własności morfologiczne. Wzorce pozbawione typowych własności morfologicznych można rozpoznać np. po tym, że szum jest przedstawiany w nielosowy sposób lub jest nielosowo rozproszony, albo po tym, że wzorzec radykalnie różni się od tych dotychczas rozpoznanych. Skuteczność przyczynowa oznacza zaś, że określone wzorce są istotne ze względu na ich konstytutywną rolę w przewodzeniu działaniu. Do tego wątku jeszcze powrócę. Oczywiście trzeba podkreślić, że ani skuteczność przyczynowa, ani typowość własności morfologicznych nie przesądzają jednoznacznie sprawy na rzecz realności danego wzorca, niemniej są one istotną wskazówką, która pozwala nam podejrzewać, że dany wzorzec nie odpowiada określonym strukturom w świecie. Szczególnie istotna w kontekście odrzucenia sceptycyzmu przez Hohwy'ego jest przyczynowa skuteczność wzorców. Mianowicie: wzorce, które odpowiadają określonym stanom rzeczy, pozwalają na skuteczne działanie podmiotu w środowisku. Wyobraźmy sobie, że idziemy przez Saharę i czując olbrzymie pragnienie, dostrzegamy oazę. Wzorzec oazy może być dla nas przyczynowo skuteczny, tzn. ma on określoną siłę dowodową, zwiększającą wierzytelność jego modelu, o ile nie jest on fatamorganą, czyli złudzeniem optycznym. Podmiot co prawda rozpoznaje taki wzorzec. Może on również mieć typowe cechy morfologiczne, ale w żaden sposób nie przyczynia się do sukcesu działania w środowisku, tym samym maksymalizuje potencjalny błąd predykcyjny.

Rola realnych wzorców, tj. takich, jakie mają siłę dowodową, czyli tych odnoszących się do określonych stanów świata, jest kluczowa dla procesu minimalizacji błędu predykcyjnego. Gdyby wzorce nie były realne, to błąd predykcyjny byłby nieustannie zwiększany, a nie redukowany. Proces rozpoznawania realnych wzorców jest ściśle powiązany z ważeniem precyzji. Wyobraźmy sobie, że szukamy solniczki w kuchni. Co kieruje naszą uwagą? Określone oczekiwania, np. to, że solniczka jest

niedużym przedmiotem, że zazwyczaj stoi na półce z przyprawami, że wypełnia ją sypki, biały proszek, który ma słony smak. Większą precyzję przypiszemy błędowi predykcyjnemu, zgodnie z którym solniczka leży na blacie stołu, aniżeli takiemu, który zakłada jej lokalizację w lodówce. Czy zatem podmiot będzie szukał wzorca, który zwiększa błąd predykcyjny, czyli takiego, jakiemu przypiszemy małą precyzję, czy raczej tego, który go zmniejsza, tj. takiego, jaki ma dużą precyzję? Oczywiście tego drugiego, bowiem wiarygodne (*reliable*) wzorce to takie, jakie mają wysoki stosunek sygnału do szumu. Ważenie precyzji w tym przypadku sprowadza się do zmiany wzmocnienia określonych błędów predykcyjnych, dzięki czemu możliwe jest elastyczne równoważenie wpływu oddolnego sygnału i generowanych odgórnie predykcji. Skuteczne działanie w środowisku (oparte na predykcjach przyszłych wydarzeń) zakłada, że model musi wyciszyć ważenie wybranych aspektów danego błędu predykcyjnego (Friston, Mattout, Kilner, 2011). Owo wyciszenie jest współkonstytuowane na podstawie ustrukturyzowania sygnału zmysłowego, w którym rozpoznajemy określone dane jako „uformowane" wzorce. Kwisthout, Bekkering, van Rooij (2017) pokazali, że zarówno zwiększenie poziomu szczegółowości modelu, jak i obniżenie poziomu szczegółowości predykcji mogą być mechanizmami, które skutecznie minimalizują błędy predykcyjne, ponieważ oba zwiększają precyzję. Wiąże się to jednak z obniżeniem ilości informacji. Innymi słowy, dla szacowania precyzji wcale nie jest konieczne, aby informacja o błędzie była bardzo szczegółowa. Zatem optymalniejsze dla podmiotu jest postrzeganie danych ustrukturyzowanych jako wzorce statystyczne aniżeli w innej, bardziej skomplikowanej czy złożonej formie. Przykładowo „bodziec wizualny to rudzik" to predykcja na dość wysokim poziomie szczegółowości, natomiast „bodziec wizualny to ptak" to predykcja na niższym poziomie szczegółowości (Kwisthout, Bekkering, van Rooij, 2017, s. 88). Zazwyczaj łatwiej jest nam zauważyć ptaka, niż dostrzec specyficzne trzepotanie skrzydłami charakterystyczne dla rudzika (Friston i in., 2012a, s. 4). Dzieje się tak, ponieważ wzorce mogą być „nieprecyzyjne". David Wallace (2003, s. 95) pokazuje, że istnieją sytuacje, w których wzór może zawierać określoną ilość szumu bądź być niedokładny, pozostając przy tym takim samym wzorem. Należy zatem powiedzieć, że w kontekście przeprowadzonych dotąd analiz realność wzorca zwiększa precyzję określonego błędu predykcyjnego, tym samym zwiększając szansę na sukces działania podmiotu w środowisku.

Wracając do przykładu solniczki: gdyby statystyczny wzorzec solniczki nie odpowiadał prawdziwej solniczce, to solniczki zapewne służyłyby innym celom, niż służą. To, że raczej wybierzemy solniczkę do posolenia zupy aniżeli pieprzniczkę, dowodzi realności wzorców, które rozpoznajemy. Dobrze można to opisać, używając kategorii działania rozumianego jako wnioskowanie aktywne: jeżeli pojawia się błąd predykcyjny dotyczący oczekiwanej i faktycznej lokalizacji solniczki, to skutecznie zminimalizujemy ten błąd, interweniując w określony stan rzeczy

i zmieniając położenie solniczki. Skuteczność aktywnego wnioskowania jest więc oparta na faktycznej skuteczności przyczynowej solniczki jako przedmiotu, który służy „do solenia".

Interesująca dla niniejszych analiz jest możliwość popełnienia błędu. To, że istnieje możliwość pomylenia się, np. potraktowania pieprzniczki jako solniczki, przemawia na korzyść tezy, iż rozpoznajemy realne wzorce. Wanja Wiese (2016b; por. także McKay, Dennett, 2009) przekonująco wykazuje, że możliwość popełniania błędu nie jest tylko słabością naszego aparatu percepcyjnego, ale wprost wynika z naszej ewolucyjnej przeszłości i ma określoną wartość adaptacyjną. Gdyby istniał złośliwy demon generujący fałszywe przyczyny naszych wzorców statystycznych, to działania nie pozwalałyby weryfikować prawdziwości przynajmniej niektórych z naszych przekonań. To, że solniczka ma określoną skuteczność przyczynową, opiera się na tym, iż pozwala ona zweryfikować przekonanie, zgodnie z którym służy ona do solenia dania[49]. Można zatem powiedzieć, że informacja o solniczce docierająca do modelu ma wartość dowodową (por. Hohwy, 2014, s. 5) w uzasadnianiu predykcji dotyczących możliwości posolenia dania. Ma ona jednak, co podkreślam, wartość dowodową, o ile jest przyczynowo skuteczna, co oznacza, że wzorcowi solniczki odpowiadają określone struktury w świecie.

Podsumowując: wykazałem, że internalizm PP nie uzasadnia przekonania Zahaviego, iż „świat doświadczania jest w sobie tylko reprezentacyjnym konstruktem" (Zahavi, 2018, s. 48). Konserwatywne PP nie ma konsekwencji antyrealistycznych i nie prowadzi do stanowiska antyrealizmu. Przeciwnie zakłada, że dane zmysłowe są ustrukturyzowane w formie określonych wzorców, które są rzeczywiste. Należy stwierdzić, że ustrukturyzowanie sygnału zmysłowego nie jest arbitralne w tym sensie, że jest tylko i wyłącznie zależne od aktywności modelu generatywnego, ale ma ono określoną skuteczność przyczynową wynikającą z tego, że przynajmniej niektóre ze wzorców w tym sygnale odpowiadają określonym strukturom w świecie. Teraz będę argumentował za twierdzeniem, że S-reprezentacje postulowane przez konserwatywne PP reprezentują realne wzorce, co oznacza, iż nie są one po prostu konstruktami mentalnymi (por. Zahavi, 2018, s. 49), którym nic nie odpowiada w świecie realnym. Zatem, wbrew Zahaviemu, treść naszego doświadczenia nie jest „generowaną przez mózg symulacją" (Zahavi, 2018, s. 7).

[49] Uzasadnienie przekonań zależy nie tylko od procesu, który je wytworzył, ale także od tego, na ile te procesy są wiarygodne w świecie rzeczywistym. W tym sensie uzasadnione przekonanie to takie, jakie – mówiąc wprost – pochodzi z efektywnych operacji poznawczych (Goldman, 1986, s. 182). Zwolennik PP może zatem twierdzić, że percepcja musi być przynajmniej w pewnym stopniu „wierna" rzeczywistości, w przeciwnym razie nie mogłaby skutecznie minimalizować błędów predykcyjnych. Minimalizując potencjalne błędy predykcyjne, model generatywny oczekuje, że informacje docierające do niego będą do pewnego stopnia wiarygodne. Gdyby tak nie było, aktualizacja jego wewnętrznego modelu świata nie byłaby możliwa. Chodzi zatem o to, aby zapobiec nadmiernemu zniekształceniu modelu przez niewiarygodne dane.

1.4.1.4. S-reprezentacje i realne wzorce

S-reprezentacje są to takie reprezentacje mentalne, jakie działają na podstawie podobieństwa strukturalnego[50] pomiędzy samą reprezentacją a tym, co reprezentowane (por. Cummins, 1996; Gładziejewski, 2016; Gładziejewski, Miłkowski, 2017; O'Brien, Opie, 2004; Shagrir, 2012; Shea, 2014). Przewodzą one działaniom swoich użytkowników, są odłączalne, a także pozwalają swoim użytkownikom na rozpoznawanie błędu reprezentacyjnego[51]. Istnieją mocne argumenty na rzecz tezy, że reprezentacje postulowane przez konserwatywne PP są strukturalne (por. Gładziejewski, 2016; Kiefer, Hohwy, 2018). Nie będę ich tu zatem szczegółowo odtwarzał. Ograniczę się do omówienia ich roli w wyjaśnianiu systemów poznawczych. Paweł Gładziejewski i Marcin Miłkowski (2017) twierdzą, że nośniki wewnętrznych S-reprezentacji można traktować jako części składowe mechanizmów poznawczych, a także jako cele różnych operacji poznawczych. Ich argumentacja opiera się na stwierdzeniu, że reprezentacje strukturalne są istotne przyczynowo (*casually relevant*). Oznacza to, że wyjaśnienia zjawisk poznawczych, które odwołują się do pojęcia S-reprezentacji, mogą być prawdziwe ze względu na podobieństwo pomiędzy reprezentacją a jej celem. Korespondencja strukturalna może dosłownie (*literally*) spowodować, że użytkownik danej reprezentacji może osiągnąć sukces w odniesieniu do celu, dla którego wykorzystał tę reprezentację. Zaś brak korespondencji może być dosłownie przyczyną porażki dla działań lub celów danego użytkownika[52] (Gładziejewski, Miłkowski, 2017, s. 340). Sukces lub porażka są zatem przyczynowo zależne od zachodzenia podobieństwa strukturalnego lub jego braku. Z tego powodu wyjaśnienia wykorzystujące S-reprezentacje powinno się traktować jako wyjaśnienia przyczynowe, w których *explanansem* są pewne fakty dotyczące podobieństwa, zaś *explanandum* sukces lub porażka danego użytkownika. Takie ujęcie jest jednak mało precyzyjne. Odpowiednią ramę interpretacyjną wyznacza mechanistyczna koncepcja wyjaśnień naukowych (por. Bechtel, 2008; Craver, 2007).

W tym modelu wyjaśnienia mechanizm, za pomocą którego należy wyjaśnić dane zjawisko, powinien być możliwy do zidentyfikowania i opisania. Oznacza to z jednej strony, że zjawisko obejmuje pewne elementy funkcjonalne, do których można przypisać odpowiednie operacje, z drugiej zaś, że mechanizm wyjaśnia-

[50] Pojęcie to Gerard O'Brien i John Opie (2004, s. 11) definiują następująco: „Suppose SV = (V, RV) is a system comprising a set V of objects, and a set RV of relations defined on the members of V. [...] We will say that there is a second-order resemblance between two systems SV = (V, RV) and SO = (O, RO) if, for at least some objects in V and some relations in RV, there is a one-to-one mapping from V to O and a one-to-one mapping from RV to RO such that when a relation in RV holds of objects in V, the corresponding relation in RO holds of the corresponding objects in O".

[51] Pełną charakterystykę S-reprezentacji przedstawiają prace: Gładziejewski i Miłkowski (2017) oraz Gładziejewski (2016).

[52] Dobrze to obrazuje prototypowy przykład S-reprezentacji, jaką jest mapa kartograficzna (por. Gładziejewski, 2016, s. 566–569). Tutaj również autor podaje uzasadnienie dla twierdzenia, że mapy można traktować jako S-reprezentacje.

jący zjawisko można podzielić na wzajemnie zintegrowane części, które są odpowiedzialne za zjawisko (Zednik, 2008, s. 1454). W tym podejściu „mechanizm" jest definiowany jako „struktura pełniąca określone funkcje na podstawie części składowych, operacji składowych i ich organizacji. Zorganizowane działanie mechanizmu odpowiada za jedno lub więcej zjawisk" (Bechtel, 2008, s. 13). Bardzo ważnym elementem wyjaśnień mechanistycznych jest tzw. dekompozycja, która polega na identyfikacji obiektów i procesów stanowiących dany mechanizm. Ogólnie mówiąc, wyjaśnić dane zjawisko, to wskazać mechanizm odpowiedzialny za jego realizację. Gładziejewski i Miłkowski wykazują, że każdy mechanizm wyposażony w S-reprezentację jako swoją część składową jest podstawą dla określonej zdolności poznawczej. Zatem S-reprezentacje interpretowane jako komponenty mechanizmu zawdzięczają swoją charakterystykę funkcjonalną sposobowi, w jaki wnoszą swój wkład do danego zjawiska, za które odpowiedzialny jest większy mechanizm. Znaczy to, że strukturalne podobieństwo pomiędzy reprezentacją a tym, co ona przedstawia, reprezentuje to, co przyczynia się do prawidłowego funkcjonowania danego mechanizmu.

Tak rozumiane reprezentacje pełnią zatem określone funkcje przyczynowe w wyjaśnianiu zjawisk. Czynią one zatem zadość „wymogowi opisu zadania" (*job description challenge*) (Ramsey, 2007). Warto zwrócić uwagę na to, że w literaturze filozoficznej zaproponowano, aby mapy przestrzenne hipokampu u szczurów potraktować jako dobry przykład wewnętrznej S-reprezentacji (por. Ramsey, 2016; Shea, 2014). Hipokamp szczura uważa się za implementację wewnętrznej mapy układu przestrzennego środowiska, zakodowanego w kartezjańskim układzie współrzędnych. Wzorce współaktywacji tzw. komórek miejsca w hipokampie mają odpowiadać strukturze przestrzennej środowiska szczurów (Shea, 2014). Oznacza to, że wzorzec relacji współaktywacji pomiędzy komórkami miejsca (w przybliżeniu tendencja poszczególnych komórek do wykazywania wspólnej aktywności) przypomina strukturę relacji metrycznych pomiędzy lokalizacjami w obrębie środowiska. W takim ujęciu mapa hipokampalna stanowi komponent mechanizmu poznawczego, na którym opiera się zdolność nawigacji po środowisku (Craver, 2007). Jest to pośredni dowód na rzecz tezy o ontycznym charakterze pojęcia S-reprezentacji, który jest istotny dla stanowiska realizmu naukowego[53]. Mianowicie: wyjaśnienie funkcji hipokampu szczura za pomocą S-reprezentacji pokazuje, że pojęcie to nie tylko ma zastosowanie instrumentalne czy epistemiczne, ale że pomiędzy tym pojęciem a mechanizmem obecnym w mózgu szczura zachodzi istotny związek. Inaczej mówiąc, pojęcie S-reprezentacji spełnia tutaj nietrywialną funkcję eksplanacyjną.

[53] W tym kontekście warto dodać, że May-Britt i Edvard Moserowie oraz John O'Keefe w 2014 roku dostali Nagrodę Nobla za odkrycie komórek miejsca, które są dosłownymi mapami poznawczymi.

Powiążmy teraz analizy S-reprezentacji z problemem rozpoznawania wzorców. Obszar badań nad rozpoznawaniem wzorców można podzielić na dwa sub-pola: podejście statystyczne i strukturalne. W pierwszym podejściu wzorce są reprezentowane przez wektory cech takie, że (*x1*, ..., *xn*) ∈ *Rn*. Proces rozpoznawania wzorców opiera się zatem na założeniu, że wzorce tej samej klasy znajdują się w zwartym obszarze *Rn*. W przypadku podejścia strukturalnego wzór jest reprezentowany za pomocą grafów. Grafy są znacznie wydajniejsze i elastyczniejsze niż wektory, ponieważ wektory cech nie dają bezpośredniej możliwości opisania relacji strukturalnych w rozpatrywanych wzorach (por. Danks, 2014). Dany element wzorca można oznaczyć jako należący do odpowiedniej klasy, jeśli jego reprezentacja (np. graficzna albo neuronalna jak w przypadku szczura) jest izomorficzna z prototypowymi grafami tej klasy. Podejście strukturalne pozwala na wykrywanie odwzorowania korespondencji pomiędzy dwoma przedmiotami, np. reprezentacją a określonym wzorcem. Główną zaletą stosowania reprezentacji strukturalnej określonych wzorców jest to, że liczba cech takiej reprezentacji nie jest stała[54], a związek pomiędzy poszczególnymi komponentami cech pojawia się wyraźnie w reprezentacji (w przeciwieństwie do reprezentacji wektora cech charakterystycznej dla podejścia statystycznego) (Bunke, Riese, 2008).

Należy zatem stwierdzić, że S-reprezentacja nie tyle odwołuje się do określonego zbioru cechy danego wzorca, ile do izomorficznej relacji pomiędzy reprezentacją a wzorcem. Jeżeli wzorce są realne, za czym argumentowałem, to ich strukturalne reprezentacje mają określone moce przyczynowe oraz eksplanacyjne ze względu na podobieństwo funkcjonalnej relacji pomiędzy zakodowanymi zmiennymi a strukturą przyczynową środowiska. Oczywiście zgodnie z PP, S-reprezentacja może reprezentować określone wzorce rozpoznawalne w sygnale zmysłowym, ale są to wzorce, które odpowiadają takim, a nie innym stanom świata. Z tego właśnie powodu sukces lub porażka są przyczynowo zależne od zachodzenia podobieństwa strukturalnego lub jego braku. Jeżeli odpowiedniemu wzorcowi nie odpowiada faktyczny stan świata, to działanie podmiotu skazane jest na porażkę[55]. Innymi słowy, skuteczność działań podmiotu jest przyczynowo zależna od strukturalnego podobieństwa pomiędzy danym wzorcem a modelem. Gładziejewski (2016, s. 570–571) trafnie opisuje ten mechanizm:

> przyczynowe – probabilistyczne prawidłowości w świecie zewnętrznym determinują wzorce statystyczne obecne na wejściach zmysłowych systemu. Statystyka nadchodzących sygnałów zmysłowych służy jako jedyny dostępny dla systemu *ślad* przyczynowo probabilistycznej struktury świata zewnętrznego. Z punktu widzenia

[54] W PP wiązałoby się to z faktem, że dla szacowania precyzji wcale nie jest konieczne, aby informacja o błędzie była bardzo szczegółowa, co wykazałem wcześniej.

[55] Nie wykluczam jednak możliwości, że model może mieć reprezentację wzorca, który nie ma siły dowodowej, czyli jest nieskuteczny przyczynowo.

> struktury kodowania predykcyjnego nasz system poznawczy wykorzystuje statystyki danych wejściowych w celu zbudowania wewnętrznego modelu świata zewnętrznego, który wytwarza takie dane wejściowe. Innymi słowy, zakładając, że istnieje funkcja, która probabilistycznie odwzorowuje stany świata (zewnętrzne przyczyny sygnału zmysłowego, tj. przedmioty z ich wzorzystymi oddziaływaniami) na stany systemu, wówczas zadaniem jest wewnętrzne odtworzenie tej funkcji poprzez zbudowanie modelu, który jest surogatem środowiska zewnętrznego.

Niniejsze analizy dowodzą, że jeżeli nawet S-reprezentacje są neurореprezentacjami, co poświadcza wiele badań (por. Ramsey, 2016; Shea, 2014), to ich wykorzystanie przez model generatywny w żadnym stopniu nie implikuje antyrealizmu, ponieważ (1) S-reprezentacje są oparte na podobieństwie strukturalnym zachodzącym pomiędzy określonymi parametrami modelu a realnymi wzorcami rozpoznawalnymi w sygnale zmysłowym; (2) są one reprezentacjami realnych wzorców, ponieważ mają określone moce przyczynowe (dzięki czemu mogą m.in. kierować skutecznymi działaniami podmiotów); oraz (3) mogą służyć jako przyczynowe wyjaśnienia określonych działań w tym sensie, że sukces lub porażka danego działania zależy od wystąpienia strukturalnego podobieństwa lub jego braku (por. Gładziejewski, Miłkowski, 2017). Oznacza to, że zarzut sformułowany przez Zahaviego, jakoby neurореprezentacjonizm w PP prowadził do rozstrzygnięć antyrealistycznych, jest fałszywy.

Kończąc dyskusję z krytyką Zahaviego, odniosę się jeszcze do problemu pośredniości percepcji.

1.4.2. Pośredniość percepcji w przetwarzaniu predykcyjnym

Zdaniem Zahaviego bezpośrednie przedmioty percepcji są w PP mentalnymi konstruktami. Oznacza to, że nigdy nie mamy bezpośredniego dostępu do przedmiotów znajdujących się w świecie. Zahavi powołuje się w tym miejscu na słowa Hohwy'ego, który twierdzi, że istnieje „granica dowodowa" (*evidentiary boundary*) oddzielająca mózg od wszystkiego, co znajduje się poza granicami czaszki (Hohwy, 2014). Takie postawienie sprawy, według Zahaviego, prowadzi do dualizmu świata i mózgu. Nie istnieje bezpośredni dostęp do rzeczywistości, ponieważ percepcja jest formą testowania hipotez (por. Gregory, 1980) i ostatecznie polega ona na wnioskowaniu o przyczynach pobudzeń zmysłowych na podstawie treści sygnału zmysłowego. Z powodu takiego rozumienia percepcji, zdaniem Zahaviego, PP staje się współczesną formą neokantyzmu wraz z jego epistemologicznymi i metafizycznymi konsekwencjami (Zahavi, 2018, s. 49–53).

Krytyka Zahaviego jest w pewnym stopniu zrozumiała. Analizując bowiem wypowiedzi Jakoba Hohwy'ego i Andy'ego Clarka, można zrozumieć, dlaczego autor

ten wprost deklaruje, że PP zakłada jedynie pośredni dostęp do rzeczywistości na zewnątrz systemu poznawczego, i z tego powodu „powinniśmy odrzucić wszystkie twierdzenia dotyczące istnienia ścisłego połączenia umysłu ze światem" (Zahavi, 2018, s. 49). Przyjrzyjmy się temu bliżej.

Hohwy stwierdza, że „percepcja jest pośrednia […] to, co widzimy jest najlepszą hipotezą mózgu" (2007, s. 322). Jest ona pośrednia w tym sensie, że zewnętrzne stany rzeczy „mogą pozostać ukryte za zasłoną danych zmysłowych i jedynie można o nich wnioskować" (Hohwy, 2013, s. 50). Ponadto autor ten (2013, s. 228) pisze:

> Zadowalającym nie jest zatem twierdzenie, że relacja percepcyjna jest bezpośrednia, ani też, iż jest ona pośrednia. Właściwą odpowiedzią na tę sytuację nie jest po prostu wymuszanie wyboru pomiędzy tymi opcjami, ale taka próba rekonstrukcji percepcyjnego związku ze światem, w jakiej nie musimy przede wszystkim wybierać pomiędzy „bezpośredniością" lub „pośredniością" spostrzeżenia.

Chodzi o to, że „mózg z wnętrza czaszki musi wnioskować o ukrytych przyczynach swoich pobudzeń zmysłowych, co jest w znacznym stopniu zależne od jego interakcji z ciałem" (Hohwy, 2013, s. 220) itd. Nietrudno zauważyć, że Hohwy z jednej strony podkreśla pośredni charakter percepcji, z drugiej zaś nie ma pełnego przekonania, czy taki opis w sposób satysfakcjonujący zdaje sprawę z tego, jaka jest relacja mózgu i środowiska, bowiem ten ostatni można traktować jako organ wyjaśniający ucieleśniony umysł (por. Hohwy, 2015a, 2018).

Przejdźmy do Clarka. Jego zdaniem problem bezpośredniości lub pośredniości percepcji jest „delikatną kwestią" (*delicate issue*), której nie będzie rozstrzygał (Clark, 2013a, s. 493). W innym miejscu twierdzi, że „trudno jest systemom zmysłowym wywodzić przyczyny pobudzeń zmysłowych z ich cielesnych efektów" (Clark, 2016, s. 19). „Mózg musi odkrywać informacje o prawdopodobnych przyczynach sygnałów zmysłowych bez żadnej formy bezpośredniego dostępu do ich źródeł" (Clark, 2016, s. 16). „Wszystko to, co 'wie', w każdym bezpośrednim znaczeniu tego słowa, to sposoby przepływu i zmiany własnych stanów […] cały system ma bezpośredni dostęp tylko do swoich własnych stanów" (Clark, 2013b, s. 183). „To, co spostrzegamy, to świat (miejmy nadzieje, że) ujawniony przez najlepszą hipotezę. Nie ma zatem sensu traktować przedmiotów percepcji jako moore'owskich 'danych zmysłowych', jeżeli są one pomyślane jako pośredniki pomiędzy spostrzegającym a światem" (Clark, 2013b, s. 492). Na koniec zaś dodaje, że PP wiąże się z metafizyczną perspektywą, którą można najtrafniej określić jako „niepośrednią percepcję" (*not-indirect perception*) (Clark, 2013b, s. 493). Clark, podobnie jak Hohwy, jest ambiwalentny w stosunku do problemu pośredniości/bezpośredniości percepcji. Co ciekawe, odrzuca on klasycznie rozumiany fenomenalizm, zakładający istnienie jakichś przedstawień czy dat wrażeniowych pomiędzy podmiotem a światem, niemniej tylko o dostępie do własnych stanów wewnętrznych mówi, że jest bezpośredni.

Drayson w ważnym artykule pt. *Direct perception and the predictive mind* (2017) przekonująco wykazuje, że zwolennicy PP używają pojęć bezpośredniości i pośredniości w trzech niesprowadzalnych do siebie znaczeniach: psychologicznym, metafizycznym i epistemologicznym. Co ważniejsze, jej zdaniem, poszczególni autorzy zajmujący się PP często nie odróżniają sensów pojęcia bezpośredniości/pośredniości, którymi się posługują, co dodatkowo utrudnia właściwe odczytanie ich intencji. W praktyce oznacza to, że przykładowo percepcja może być pośrednia w sensie psychologicznym, zaś bezpośrednia w sensach metafizycznym i epistemologicznym. Przyjrzyjmy się bliżej poszczególnym znaczeniom.

Mówiąc o psychologicznym znaczeniu pojęcia bezpośredniości/pośredniości, ma się na myśli to, że percepcja jest procesem inferencyjnym bądź nim nie jest. Źródeł tego odróżnienia należy szukać w pracach Jamesa J. Gibsona (1967, 1979), który pokazał, że pytanie o psychologiczną bezpośredniość percepcji dotyczy przede wszystkim sposobu, w jaki bodźce (informacje docierające przez wejścia zmysłowe) są przetwarzane. Tym sposobem Gibson przeciwstawił ujęcia rozumiejące percepcje konstruktywistycznie, czyli na podstawie tzw. wnioskowania percepcyjnego (por. Gregory, 1980), i wewnętrzne reprezentacje ujęciu ekologicznemu, które zakłada „bezpośrednią lub natychmiastową świadomość (*awareness*) obiektów oraz zdarzeń", które są percypowane (Gibson, 1967, s. 168). Ujęcie konstruktywistyczne zakłada pośredniość psychologiczną percepcji, zaś ekologiczne jej psychologiczną bezpośredniość.

Metafizyczne pojęcie bezpośredniości/pośredniości dotyczy sposobu, w jaki należy myśleć o przedmiotach percepcji oraz relacji pomiędzy tymi przedmiotami a podmiotem je postrzegającym. Sposób ten powinien oddawać sprawiedliwość naszym doświadczeniom weredycznym, czyli takim, o jakich możemy orzekać prawdziwość bądź fałszywość, jednocześnie uwzględniając nasze doświadczenia nieweredyczne (np. sny) (por. Burge, 2010; Drayson, 2017, s. 9). Reakcją na ten problem są takie metafizyczne teorie percepcji jak naiwny realizm, fenomenalizm czy teoria danych zmysłowych. Przykładowo naiwny realista, na którego powołuje się Zahavi, utrzymuje, że sam świat jest przedmiotem jego percepcji. Prawdziwym doświadczeniem jest percepcja przedmiotów fizycznych i ich własności. W doświadczeniu weredycznym te przedmioty wraz ze swoimi własnościami stanowią o jego treści. Zaś w doświadczeniu nieweredycznym nie są one tym, czego się doświadcza, zatem nie mogą stanowić jego treści (sny są przykładową treścią takich doświadczeń). Prowadzi to do konkluzji, która zakłada tzw. dysjunktywizm metafizyczny, zgodnie z którym istnieje zarówno percepcja weredyczna, jak i nieweredyczna, nawet jeżeli dana osoba nie potrafi wykazać różnicy pomiędzy tymi rodzajami doświadczeń (por. Burge, 2005; McDowell, 2013). Aby zachować swoje przekonania o metafizycznej bezpośredniości percepcji, zwolennik tego ujęcia jest zobowiązany przyjąć nieintuicyjne twierdzenie o metafizycznej indywiduacji swoich

stanów percepcyjnych (Drayson, 2017, s. 10). Zwolennik teorii danych zmysłowych będzie utrzymywał, że percepcja zasadniczo opiera się na spostrzeganiu wewnętrznych bytów umysłowych lub „danych zmysłowych". Weredyczne i nieweredyczne przypadki doświadczenia percepcyjnego można zatem zaklasyfikować jako ten sam rodzaj doświadczenia ze względu na ten sam rodzaj przedmiotu odniesienia. Percepcja jest zatem weredyczna wtedy, gdy istnieje korespondencja pomiędzy daną zmysłową a światem zewnętrznym, nieweredyczna jest zaś wtedy, gdy takiej korespondencji nie ma.

Bezpośredniość/pośredniość w sensie epistemologicznym odnosi się do charakteru naszych przekonań percepcyjnych na temat świata oraz sposobu ich uzasadniania. Jeżeli przekonania są uzasadniane nieinferencyjnie, są one bezpośrednie. Jeżeli zaś uzasadnia się je w sposób inferencyjny, to są one pośrednimi przekonaniami percepcyjnymi. Dane przekonanie jest uzasadnione nieinferencyjnie wtedy, gdy jest uzasadniane przez takie stany, jakie same nie podlegają uzasadnieniu. Inaczej mówiąc, przekonanie jest bezpośrednio uzasadnione, jeżeli nie uzasadnia się go na podstawie innego przekonania, ale np. przez konkretne doświadczenie percepcyjne. W tym ujęciu naiwny realista zakłada bezpośrednie uzasadnienia przekonań percepcyjnych, zaś teoretyk danych zmysłowych pośrednie. Zdaniem Drayson problemy, które dotyczą epistemologicznego pojęcia bezpośredniości/pośredniości percepcji w PP, biorą się z różnych funkcji, jakie inferencjom przypisują badacze. Raz są one wykorzystywane jako kroki w rozumowaniach, a innym razem jako elementy uzasadniające dane twierdzenia (Drayson, 2017, s. 16). Należy zauważyć, że zarówno Hohwy, jak i Clark od psychologicznej pośredniości percepcji w PP wprost przechodzą do tezy o ich epistemologicznej pośredniości, co zdaniem Drayson jest nadużyciem.

Według tej autorki w PP percepcję należy rozumieć jako pośrednią w sensie psychologicznym, co oznacza, że zakłada się tutaj, iż każda informacja musi być w jakiś sposób przetworzona, zanim zostanie przez system poznawczy wykorzystana. Nie przeważa to jednak szali na rzecz twierdzenia, że percepcja jest pośrednia w pozostałych dwóch sensach, tj. metafizycznym i epistemologicznym. Zdaniem Drayson PP pozwala zarówno na interpretację percepcji w kategoriach metafizycznej i epistemologicznej bezpośredniości lub pośredniości. Co oznacza, że problem ten nie jest *a priori* przesądzony, jak mógłby sugerować Zahavi. Powyższe analizy mają więc kapitalne znaczenie dla odpowiedzi na zarzuty i krytykę sformułowaną przez Duńczyka i innych badaczy. Mianowicie: nie jest zupełnie jasne, jakie znaczenie Zahavi nadaje pojęciom bezpośredniości i realizmu. Broniona przez niego postawa, zakładająca „naszą naturalną skłonność do realizmu", nie pozwala na uzasadnienie twierdzenia, że PP jest z gruntu konstruktywistyczne i antyrealistyczne, ponieważ zdaniem Drayson tylko w sensie psychologicznym można orzekać, że percepcja w PP jest pośrednia. Niemniej pośredniość psychologiczna nie wyklucza ani

stanowiska realizmu metafizycznego czy epistemologicznego. Krytyki Zahaviego w takim ujęciu nie można traktować jako rzeczowej, ponieważ nie wiadomo tak naprawdę, z jakim stanowiskiem i jaką jego wykładnią autor ten polemizuje.

Zahavi wydaje się także, że nie odróżnia pośredniości percepcji od jej systematycznej niereliabilności. Chodzi o to, że dane przekonanie jest uzasadnione wtedy i tylko wtedy, gdy zostało wytworzone przez rzetelny proces poznawczy. Percepcja zatem, aby mogła być rzetelnym procesem poznawczym, musi prowadzić do uzasadnionych przekonań, które mogą być punktem wyjścia do przeprowadzania wnioskowań, a także do powstania nowych uzasadnionych przekonań, co w efekcie prowadzi do działań, które są skuteczne w określonym środowisku. Istotne jest to, że uzasadnienie przekonania zależy nie tylko od procesu, który go wytworzył, ale również od tego, czy procesy te są rzetelne w świecie realnym. W tym sensie uzasadnione przekonanie to takie, jakie – w dużym uproszczeniu – pochodzi z operacji poznawczych, które są skuteczne (Goldman, 1986, s. 182). Zwolennik PP, co wykazałem, może zatem twierdzić, że percepcja jest w jakimś sensie zapośredniczona, jednocześnie broniąc tezy, że tak rozumiana percepcja musi być chociaż w jakimś stopniu „wierna" rzeczywistości, inaczej bowiem nie mogłaby skutecznie minimalizować błędu predykcyjnego. Model generatywny, minimalizując potencjalne błędy predykcyjne, oczekuje, że informacje docierające do mózgu są w jakimś stopniu wiarygodne (*reliable*). Inaczej nie mógłby modyfikować swojego modelu świata. Chodzi zatem o to, żeby zapobiec nadmiernemu zniekształcaniu modelu przez niewiarygodne dane. Oznacza to, że pośredni charakter percepcji nie implikuje, jak sugeruje Zahavi, braku odniesienia do świata. Oczywiście istnieją operacje poznawcze, które mają większy stopień uzasadniania prawdziwości przekonań (np. długotrwała obserwacja), oraz te, które uzasadniają prawdziwość przekonań w stopniu słabszym (np. pobieżne spostrzeżenie). Nie zmienia to jednak faktu, że w PP percepcja spełnia nie tylko funkcję dostarczania informacji, ale także jej wstępnego uzasadniania. Wiąże się to bezpośrednio z bayesowskim charakterem zmieniania i aktualizowania wewnętrznych parametrów modelu generatywnego na podstawie probabilistycznej sieci przekonań i wnioskowania abdukcyjnego (por. Kiefer, 2017).

Podsumowując, należy powiedzieć, że niniejsze analizy pozwalają, jak sądzę, odrzucić zarzut Zahaviego, jakoby PP było antyrealistyczne. Udało się bowiem wykazać, że stanowisko internalistyczne zakładające, iż granicą systemu poznawczego jest czaszka bądź granica systemu nerwowego, opiera się na ustaleniu, gdzie ma być umieszczony koc Markova, którego lokalizacja jest niezależna od jakichkolwiek faktów na temat mózgu, świata itd. To, że w konserwatywnym PP twierdzi się, iż koc najlepiej umieścić wokół mózgu, jest ustaleniem warunkowym podyktowanym m.in. strategią eksplanacyjną, zainteresowaniami badawczymi, jak i samym badanym zjawiskiem. Dodatkowo charakter i treść parametrów modelu

są wytwarzane w ścisłym związku przyczynowym ze światem, co w przypadku PP oznacza, iż ukryte przyczyny wzorców danych sensorycznych muszą zostać ujawnione we wnioskowaniu przyczynowym. Jest to możliwe dzięki temu, że przynajmniej niektóre wzorce statystyczne w sygnale zmysłowym mają określoną siłę dowodową, ponieważ zwiększają wierzytelność modelu, a zwiększają wierzytelność modelu, ponieważ odpowiadają im jakieś struktury w świecie. W tym sensie są to wzorce realne, a za ich pośrednictwem model ma dostęp do określonych stanów świata. Model generatywny wykorzystuje S-reprezentacje określonych wzorców, które (1) są oparte na podobieństwie strukturalnym zachodzącym pomiędzy określonymi parametrami modelu a realnymi wzorcami rozpoznawalnymi w sygnale zmysłowym; (2) są one reprezentacjami realnych wzorców, ponieważ mają określone moce przyczynowe (dzięki czemu mogą m.in. kierować skutecznymi działaniami agentów); oraz (3) mogą one służyć jako przyczynowe wyjaśnienia określonych działań w tym sensie, że sukces lub porażka danego działania zależy od wystąpienia strukturalnego podobieństwa lub jego braku. Zatem jeżeli nawet, jak twierdzi część badaczy, S-reprezentacje są neuroreprezentacjami, ich istnienie nie prowadzi do rozstrzygnięć antyrealistycznych, co sugeruje Zahavi[56].

Przejdę teraz do omówienia radykalnych wersji PP, pokazując, że są możliwe inne, niekonstruktywistyczne i nieinternalistyczne, sposoby interpretowania tej ramy badawczej.

1.5. Radykalne przetwarzanie predykcyjne

Koncepcje, które można włączyć w radykalne ujęcie PP, są dużo bardziej zróżnicowane niż propozycje spod znaku podejścia konserwatywnego. Można powiedzieć, że to, co jest im wspólne, to krytyka i (lub) odrzucenie jednego albo więcej zobowiązań charakteryzujących konserwatywne PP. Do ujęć radykalnych można włączyć m.in. prace takich autorów jak Andy Clark (2015a, 2016), Nico Orlandi (2016, 2017; Orlandi, Lee, 2018), Jelle Bruineberg, Julian Kiverstein czy Erik Rietveld (Bruineberg, Rietveld, 2014; Bruineberg, Kiverstein, Rietveld, 2016; Bruineberg, 2017) oraz Michael Kirchhoff (2018; Kirchhoff, Robertson, 2018). Trudno wskazać zbiór wspólnych przekonań, które podzielaliby wszyscy wymienieni autorzy. Niemniej, jak sądzę, można wskazać dwie tezy, co do których są oni raczej zgodni. Po pierwsze podkreślają ścisłą i dynamiczną zależność pomiędzy percepcją, działaniem a środowiskiem. Po drugie, co ważniejsze w tym kontekście, twierdzą, że tylko niektóre poziomy modelu generatywnego są bezpośrednio

[56] Otwarte pozostawiam pytanie dotyczące rodzaju realizmu, jaki chcielibyśmy przypisać PP. Kwestię tę przesuwam na margines moich analiz.

reprezentacyjne, inne zaś tylko pośrednio[57], ponieważ, co podkreśla Clark, są one powiązane ze światem w sposób enaktywny, co oznacza, iż celem reprezentacji jest raczej „zaangażowanie w świecie, niż przedstawienie go w neutralny sposób" (Clark, 2015b, s. 4).

Dla ujęć radykalnych istotne są inspiracje płynące z jednej strony od psychologów ekologicznych rozwijających idee zaproponowane przez Gibsona (1966, 1979), z drugiej zaś od enaktywistów korzystających z prac m.in. Francisco Vareli i Humberto Maturany (Varela, Maturana, Uribe, 1974; Varela, Thompson, Rosch, 1992).

Podejścia ekologiczne trafnie charakteryzują Rolf Pfeifer i Josh Bongard, używając do tego tzw. zasady równowagi ekologicznej. Głosi ona, że po pierwsze z perspektywy środowiska „musi istnieć zgodność pomiędzy złożonością systemów zmysłowego, motorycznego i neuronowego danego podmiotu; po drugie zaś, że istnieje pewna równowaga lub podział zadań pomiędzy morfologią [podmiotu], jego działaniami, a środowiskiem" (2006, s. 123). Oprócz postulowanej równowagi pomiędzy działającym podmiotem a jego środowiskiem istotna jest prymarna rola zmysłów w działaniu.

1.5.1. Psychologia ekologiczna

Za właściwego twórcę psychologii ekologicznej uważa się Gibsona (por. 1966, 1979). Zaproponował on nowe podejście w psychologii, którego właściwym przedmiotem badań jest zwierzę wraz ze środowiskiem, w którym ono funkcjonuje (1979, s. 1). Zwierzę i środowisko tworzą unikalną relację, której badaniem ma się zająć nauka. Autor *Ecological approach to perception* krytykował ujęcia, w których oddziela się percepcję od działania, bowiem, jak twierdził, w najbardziej podstawowej formie percepcja jest pewną formą działania. Oddzielenie percepcji od działania prowadzi do odseparowania jej od środowiska, w którym jest realizowana. Standardowe podejścia do percepcji koncentrują się zatem na pewnym konstrukcie teoretycznym, a nie na specyficznej relacji, jaką tworzy doznające zwierzę ze swoim środowiskiem. Gibson, krytykując to podejście, pokazał, że sytuacje niespotykane w normalnym życiu traktuje ono jako wzorcowe dla percepcji. Przypadki nietypowe (np. izolowanie bodźców czy unieruchomienie obserwatora) traktowane są jako paradygmatyczne. Jego zdaniem takie podejście badawcze bierze się z błędnej metodologii, dla której paradygmatyczna jest metodologia obecna w fizyce (Gibson, 1966, s. 2). W praktyce, zdaniem Gibsona, oznaczało to, że właściwym przedmiotem badań psychologów staje się mózg, co oczywiście nie jest błędem, ale istotnym zawężeniem perspektywy badawczej. Z tego też powodu tradycyjna

[57] Część badaczy należących do tej grupy odrzuca istnienie reprezentacji w ogóle (por. Kirchhoff, Robertson, 2018; Orlandi 2016, 2017).

psychologia traktuje percepcję jako pasywne odbieranie bodźców zmysłowych, z których następnie tworzone są percepty itd. (Gibson, 1966). Mówiąc precyzyjniej, standardowe podejścia zakładają, że percepcja ma charakter pasywny i jest stymulowana przez oddolnie odbierane bodźce. Reprezentacje są budowane oddolnie z danych zmysłowych, od najprostszych własności, aż po coraz bardziej złożone kompleksy. Działania są reakcją na informacje pochodzące z wejść zmysłowych. System percepcyjny zbiera te informacje w celu utworzenia wewnętrznych reprezentacji przedmiotów istniejących w świecie zewnętrznym. Następnie wykorzystuje uzyskane informacje, reprezentacje oraz zapamiętaną wiedzę do tworzenia sądów i podejmowania decyzji dotyczących działań[58] (por. także Burr, 2017; Cisek, Kalaska, 2010; Clark, 2015b).

Gibson nie zgadzał się z twierdzeniem, że zwierzę odbiera proste bodźce (na wzór tych tworzonych w laboratorium). Z tego też powodu organizmy nie wykształciły systemów do ich odbierania[59]. W zamian za to potrafią szybko i sprawnie reagować na bodźce bardziej złożone, o ile są one ważne z punktu widzenia przeżycia w danym środowisku (Klawiter, 2012, s. 263). Organizmy jednak nie odbierają tych bodźców w sposób bierny i pasywny, ale aktywnie poszukują w otoczeniu takich pobudzeń, jakie pozwalają im przetrwać i radzić sobie ze środowiskiem i zagrożeniami zeń płynącymi. Zwierzę nie jest biernym obserwatorem, ale aktywnym eksploratorem pozyskującym złożone i ustrukturyzowane informacje ze swojego otoczenia.

Przyjmując te założenia, należy stwierdzić, że właściwym przedmiotem percepcji nie są własności przedmiotów takie jak ciężar, długość czy wysokość danego przedmiotu, ale takie własności, na jakie są ukierunkowane określone możliwości działań. Na przykład, jeżeli chce się ominąć wystającą z ziemi przeszkodę, to jej wysokość albo szerokość nie jest dla nas istotna. Istotne jest to, czy damy radę ją ominąć albo przeskoczyć. Aby tego dokonać, nie musimy znać dokładnych wymiarów tej przeszkody[60]. Ważna jest informacja, którą oferuje środowisko. W związku

[58] Z takim modelem percepcji koresponduje krytyka przedstawiona przez Susan Hurley (1998, 2001) i tzw. klasyczny model kanapkowy (*classic sandwich model*), który miał być obecny w klasycznej neuronauce i kognitywistyce. Model ten zakłada, że (1) percepcja i działanie są od siebie niezależne, są procesami peryferyjnymi; (2) poznanie (myślenie) jest podstawą umysłu. Ten jest podzielony na odpowiednie moduły: poznanie zapośrednicza relację pomiędzy percepcją a działaniem, „Umysł jest kanapką, zaś poznanie jest jej wypełnieniem" (Hurley, 2001, s. 3); (3) poznanie jest centralnym zadaniem umysłu i jest ono niezależne od obwodowych procesów sensomotorycznych.

Wydaje się jednak, że takie rozumienie tzw. klasycznej kognitywistyki jest pewnego rodzaju uproszczeniem i idealizacją, zaś krytyka wczesnych modeli niezasadna z tego punktu widzenia. Już w 1976 roku Neisser krytykował takie myślenie i zaproponował cykliczny model poznania (Neisser, 1976). Por. także Miller, Galanter, Pribram, 1980.

[59] Widoczne są tutaj wyraźne inspiracje płynące z badań prowadzonych przez psychologię postaci.

[60] Pouczające w tym kontekście są badania nad umiejętnością wchodzenia na schody przeprowadzone przez Williama H. Warrena (1984). Wykazał on, że badani postrzegali możliwość wejścia na schody jako miarę ich własnej zdolności do wykonania kroku. W szczególności udało się pokazać, że najwyższy stopień, na jaki mogą wejść, wynosi 0,87 długości ich stopy. Oznacza to, że badani postrzegali odpowiednie aspekty środowiska w odniesieniu do skali własnego ciała.

z tym Gibson proponuje, „że to, co widzimy, kiedy spoglądamy na przedmioty, to ich afordancje, a nie ich własności" (Gibson, 1979, s. 134).

Afordancja jest centralnym pojęciem ekologicznej psychologii. Neologizm ten pochodzi od angielskiego czasownika *to afford*, który oznacza *oferować*, *dostarczać* czy *użyczać czegoś*. Gibson rozumiał je następująco: „Afordancje środowiska są tym, co oferuje ono zwierzęciu, tym, co coś dostarcza albo wyposaża je [...]. Jest ona [afordancja] w równym stopniu faktem należącym do dziedziny środowiska, jak i faktem z zakresu zachowania [...]. Zakłada komplementarność zwierzęcia i środowiska" (Gibson, 1979, s. 127, 129). Przykładowo: zwierzę, widząc drzewo, dostrzega w nim afordancję lub mówiąc inaczej – ofertę do wspięcia się na nie. Pień drzewa dla jednego może być „czymś, na czym można usiąść", zaś dla innego – „czymś, z czego ma się lepszą widoczność". Sam Gibson nie dość chętnie wypowiadał się na temat statusu ontologicznego afordancji, określając je zagadkowo jako własności, które nie są ani subiektywne, ani obiektywne (Gibson, 1979, s. 129). Z tego też powodu w literaturze przedmiotu istnieje spór dotyczący ich ontologicznej charakterystyki. Chemero (2003) podkreśla, że pojęcia afordancji nie należy używać analogicznie do sposobu, w jaki posługujemy się takimi pojęciami jak: przedmiot, własność czy cecha. Afordancje bowiem nie są czymś, co mogą zbadać nauki pozytywne, takie jak biologia czy fizyka. Wiąże się to ściśle z ich relacyjnym charakterem i specyficznym sposobem powiązania ze środowiskiem.

Gibson powiada, że aby spostrzec afordancję w środowisku, nie jest potrzebna żadna wewnętrzna reprezentacja. Przekonanie to wiąże się z tezą, że dostrzeżenie afordancji polega na bezpośrednim nabyciu konkretnej informacji, iż np. na coś można się wspiąć, na czymś można usiąść albo użyć czegoś jako punktu widokowego. Zwierzę, widząc przedmioty, właściwie dostrzega oferty możliwego ich zastosowania, w zależności od stanu środowiska, jak i potrzeb danego stworzenia. Afordancje rozumiane jako relacje pełnią kluczową funkcję w kontroli działań i możliwości określania wydajności odpowiednich zachowań w określonych warunkach. Pomagają one zwierzęciu skutecznie reagować, działać i poruszać się w dynamicznie zmieniającym się otoczeniu. Nie należy zatem spostrzeżenia danej afordancji wiązać z nabywaniem wiedzy w sensie propozycjonalnym. Chodzi tutaj raczej o rozwijanie swoich umiejętności na podstawie tego, co oferuje otoczenie. Znika tutaj hierarchiczny podział na wyższe i niższe procesy mentalne i zostaje zastąpiony holistyczną oraz ekologiczną teorią percepcji oraz wiedzy zanurzonej w działaniu i świecie (Schetz, 2014, s. 96–99). Jak nietrudno spostrzec teza ekologiczna wzmacnia i ugruntowuje tezę antyreprezentacjonistyczną.

Tradycyjny gibsonizm jest antyreprezentacyjny. Jest to istotne spostrzeżenie, ponieważ zwolennicy tezy ekologicznej w ramach badań nad przetwarzaniem predykcyjnym głoszą albo wprost (por. Kirchhoff, 2018; Orlandi, 2016, 2017), albo *implicite* (por. Bruineberg, Kiverstein, Rietveld, 2014; Bruineberg, Rietveld, 2016)

poglądy antyreprezentacjonistyczne. Należy jednak zwrócić uwagę na to, że po pierwsze istnieje tradycja badań nad afordancjami, niewykluczająca z góry istnienia ważnych eksplanacyjnie reprezentacji (por. Bickhard, Richie, 1983; Norman, 2018), po drugie zaś ekologiczne podejście w ramach PP nie przekreśla jednoznacznie znaczenia reprezentacji dla analizy procesów percepcyjnych czy decyzyjnych.

1.5.2. Enaktywizm

Słowo „enaktywizm" należy do jednego z najczęściej używanych w pracach filozofów, psychologów i kognitywistów[61]. Pojawia się ono w kontekście dyskusji nad percepcją, poznaniem społecznym, reprezentacjami, ucieleśnieniem czy osadzeniem w środowisku. Nie jest łatwo jasno zdefiniować i określić, czym jest enaktywizm, z powodu wielu jego odmian, jak i braku wyraźnych deklaracji pojęciowych ze strony jego zwolenników. Sam termin pochodzi od angielskiego *enaction*, dla którego trudno znaleźć polski odpowiednik[62]. Termin ten John Protevi rozumie następująco: „sposób, w jaki podmiot percepcji kreatywnie dopasowuje swoje działania do wymogów danej sytuacji" (Protevi, 2006). To, co jest wspólne różnej maści enaktywistom, to przekonanie, że aktywność poznawcza jest procesem zachodzącym w wyniku aktywnego oddziaływania autonomicznego i samoorganizującego się systemu poznawczego (organizmu) ze środowiskiem (Ward, Silverman, Villalobos, 2017, s. 4). Podejście to opiera się na dwóch podstawowych przesłankach: (1) percepcja jest działaniem; oraz (2) struktury poznawcze wyłaniają się (*emerge from*) z powtarzających się wzorców sensomotorycznych, które umożliwiają percepcyjne działania (Varela, Thompson, Rosch, 1991, s. 173). Shaun Gallagher (2017) wyróżnia następujące przekonania wspólne dla różnych odmian enaktywizmu: (1) poznanie nie jest po prostu aktywnością mózgu; (2) świat nie jest uprzednio określony czy zdefiniowany, ale jest strukturyzowany przez poznanie i działanie; (3) procesy poznawcze mają znaczenie tylko dzięki funkcji, jaką pełnią w działaniach, a nie poprzez odwzorowanie reprezentacyjne albo posiadanie wewnętrznego i powtarzalnego modelu świata; (4) istotne są związki z teorią systemów dynamicznych, podkreślającą znaczenie sprzężeń zwrotnych na linii ciało – środowisko; (5) procesy poznawcze wyższego rzędu (wyobraźnia, myślenie abstrakcyjne, deliberacja itd.) mają charakter operacyjnej *wiedzy-jak* oraz powiązane są z ucieleśnionymi i usytuowanymi działaniami.

[61] Google Scholar wymienia od roku 2014 aż 4810 prac, w których to pojęcie się pojawia (stan na 24.09.2020).

[62] Słownikowo temu pojęciu odpowiadają w języku polskim takie terminy jak: „uchwalić", „uchwalać", „stanowić", „wprowadzać", „wcielać w życie" czy „ustanawiać". Widoczne jest, co istotne – aktywny charakter tego czasownika.

W literaturze wyróżnia się następujące rodzaje podejść enaktywistycznych[63]:

1. Enaktywizm klasyczny (autopojetyczny): ujęcie to wiąże się z propozycją teoretyczną przedstawioną przez Francisca J. Varelę, Evana Thompsona oraz Eleanor M. Rosch w pracy pt. *The embodied mind: cognitive science and human experience* (1991). Zaproponowali oni, aby koncepcję umysłu ucieleśnionego oprzeć na pojęciu *autopoiesis*[64] wprowadzonym przez Humberto Maturanę i Francisca J. Varelę (1980). Kluczowe jest tutaj podkreślenie autonomii dynamicznie rozumianych systemów poznawczych oraz tzw. biologicznej ciągłości życia i umysłu. Jest to pogląd, według którego struktury i zasady charakteryzujące umysł są po prostu wzbogaconą oraz bardziej rozbudowaną wersją struktur i zasad, które ugruntowują życie.

2. Enaktywizm sensomotoryczny: ujęcie to raczej koncentruje się na wyjaśnieniu intencjonalnej natury doświadczenia percepcyjnego niż na budowaniu jakiejś jednolitej wizji umysłu. Percepcja jest tutaj rozumiana jako specyficzny rodzaj działania w środowisku, polegający na aktywnej eksploracji otoczenia (por. Noë, 2004; O'Regan, Noë, 2001).

3. Enaktywizm partycypacyjno-sensotwórczy: podkreśla się tutaj znaczenie relacji zachodzącej pomiędzy podmiotami działającymi a środowiskiem. Istotna jest sfera społeczna i interakcje międzyosobowe. Według zwolenników tego podejścia (De Jaegher, Di Paolo, 2007; Di Paolo, De Jaegher, 2012) samoorganizacja dotyczy przede wszystkim intersubiektywnych relacji, a nie pojedynczych organizmów.

4. Enaktywizm radykalny: ujęcie, które odmawia wartości eksplanacyjnej wyjaśnieniom reprezentacyjnym i opartym na pojęciu treści. Reprezentacje istnieją jedynie na poziomie języka. Treść ma charakter wyłącznie kulturowy (Hutto, Myin, 2013, 2017).

Nietrudno zauważyć, że przyjęcie jakiejkolwiek odmiany enaktywizmu ściśle wiąże się z przyjęciem stanowiska antyreprezentacjonistycznego, które opiera się na przekonaniu, iż reprezentacje mentalne nie istnieją, a jeżeli nawet istnieją, to wyjaśnienie wykorzystujące to pojęcie, nie ma wartości eksplanacyjnej. Ów reprezentacjonizm jest również kluczowy dla tych podejść do PP, które mniej lub bardziej wprost odwołują się do stanowiska enaktywistycznego. Wątkiem tym zajmuję się w rozdziale *Zasada energii swobodnej w przetwarzaniu predykcyjnym*.

Tak jak wspominałem, istotne dla enaktywizmu są pojęcia autopojezy oraz samoorganizacji. Zdaniem Maturany i Vareli charakterystyczna dla wszelkich organizmów, czy właściwiej, układów żywych jest tendencja do samoorganizacji. System żywy to system autopojetyczny. Co to znaczy? System ten nieustannie urzeczywistnia sieć procesów, które utrzymują go przy istnieniu. Kluczowe jest to, że procesy te cechują się domknięciem operacyjnym, tzn. że ich funkcjonowanie nie jest zależne od otoczenia, ale wyznaczane jest przez własną dynamikę oraz elementy składające się na dany proces. Otoczenie może wpływać na system jedynie energe-

[63] Korzystam tutaj z prac: Miłkowski, 2018a; Schetz, 2014; Ward, Silverman, Villalobos, 2017.
[64] Por. także podrozdział *Ujęcie enaktywiczne*.

tycznie, nigdy zaś informacyjnie. Działanie, na które kładą taki nacisk zwolennicy enaktywizmu, jest koniecznym wytworem procesów, które mają na celu utrzymanie układu żywego przy życiu. Dla Vareli (1997) żywa komórka jest wzorcowym przykładem systemu autopojetycznego. Powiada on, że bakteria jest oddzielona od swojego molekularnego otoczenia przez półprzepuszczalną błonę, która jest jednocześnie tworzona i utrzymywana przez sieć procesów (absorbujących różne składniki odżywcze czy wydalających odpady), które wyznaczają granicę pomiędzy żywym organizmem a środowiskiem. Komórka zatem określa własną organizację poprzez działanie systemu wytwarzającego własne komponenty. Innymi słowy, taka, a nie inna (głównie metaboliczna) organizacja komórki sprawia, że ona żyje.

Zwolennicy enaktywizmu podkreślają jego rewolucyjny charakter dla nauk poznawczych (por. np. Müller, Müller, 2007; Stewart, Gapenne, Di Paolo, 2010). Owa rewolucyjność miałaby się wiązać z odkryciem fundamentalnego znaczenia procesów autopojetycznych. Zrozumienie tego, czym jest poznanie, wymaga wyjaśnienia natury biologicznej samoorganizacji. Krytycy jednak zwracają uwagę na kilka ograniczeń tego podejścia. Miłkowski (2018a) podkreśla, że z powodu braku jasno określonych metod badawczych, konkretnych założeń teoretycznych, a także wyników eksperymentalnych nie sposób mówić o jakimś dużym czy wręcz rewolucyjnym oddziaływaniu enaktywizmu na kognitywistykę. Dodaje przy tym, że enaktywizm nie stanowi jasnej teorii procesów poznawczych, a jedynie określoną tradycję badawczą (w rozumieniu Larry'ego Laudana[65]) myślenia o pewnych problemach. Nie jest zatem jednolitą teorią naukową. Mario Villalobos (2013) wykazuje, że enaktywizm nie był w momencie swojego powstania zupełnie nowatorską koncepcją, ponieważ mocno czerpał z dorobku cybernetyki. Z tego też powodu powinniśmy o nim myśleć nie jak o podejściu rewolucyjnym, ale raczej konserwatywno-rewizjonistycznym. Inni badacze podkreślają problemy dotyczące wewnętrznych rozstrzygnięć w obrębie enaktywizmu. Przykładowo Kenneth Aizawa pokazał (2014), że pojęcie działania używane w pracach enaktywistów odbiega od jego tradycyjnego użycia w kognitywistyce. Nie byłoby w tym nic złego, powiada Aizawa, gdyby było to inne, alternatywne zastosowanie tego pojęcia. A tymczasem pojęcie działania użyte zostaje do opisu pewnego rodzaju zachowań. Tym sposobem pojęcia myślenia i poznania zostają utożsamione, co prowadzi do pomieszania tego, co wyjaśniane (*explanandum*), z tym, co wyjaśnia (*explanans*).

[65] Tradycja badawcza, powiada Laudan, to „zbiór ogólnych założeń dotyczących bytów i procesów w pewnej dziedzinie badań i poprawnych metod, jakimi należy się posługiwać, by rozważać problemy i konstruować teorie" (1977, s. 81). Tradycja badawcza, w przeciwieństwie do teorii, określa jedynie dziedzinę zastosowań należących do teorii; nie wyjaśnia i nie przewiduje zjawisk; pełni przede wszystkim funkcje heurystyczne (Sady, 2013, s. 391–392).

1.5.3. Psychologia ekologiczna, enaktywizm a przetwarzanie predykcyjne

Pomimo wielu słabości i kontrowersji wskazywanych przez różnych krytyków tradycja enaktywistyczna ma wielu zwolenników także wśród badaczy zajmujących się PP. Z powodu swojego programowego antyreprezentacjonizmu często jest kojarzona i łączona z psychologią ekologiczną (por. np. Bruineberg, Kiverstein, Rietveld, 2014; Bruineberg, Rietveld, 2016). Należy jednak zachować ostrożność i pamiętać o różnicach dzielących oba te podejścia. Varela, Thompson i Rosch (1991, s. 203–204) zwracają uwagę na to, że psychologia gibsonowska rozumie strukturę środowiska jako złożoną z obiektywnie istniejących własności, na które organizm musi reagować. Tymczasem enaktywizm kładzie nacisk na ścisłą interakcję zachodzącą pomiędzy organizmem i jego środowiskiem. Ujęcie tej relacji ma dużo dynamiczniejszy i bardziej sprzężony charakter, niż ma to miejsce w tradycji ekologicznej. Głos autorów *The embodied mind* jest istotny, aczkolwiek wydaje się, że obecnie ma on głównie wartość historyczną. Wystarczy zwrócić uwagę na późniejszy rozwój obu tradycji, żeby zauważyć, że wskazana przez nich różnica staje się coraz mniej wyraźna. Dobrym przykładem jest psychologia ekologiczna uprawiana w wersji Anthony'ego Chemero (2003, 2009, 2014). Autor ten podkreśla dynamiczny charakter relacji organizm – środowisko i stara się go opisać przy użyciu języka systemów dynamicznych. Z innej strony mamy propozycję teoretyczną radykalnego enaktywizmu Daniela D. Hutto i Erika Myina (2013), w której zupełnie rezygnuje się z biologicznej perspektywy i języka dynamicyzmu na rzecz rozważań inspirowanych pracami Donalda Davidsona czy właśnie wspomnianego Anthony'ego Chemero[66]. Analiza różnic dzielących enaktywizm i psychologię gibsonowską wykracza poza rozważania przyjęte w tej pracy, ale muszę wspomnieć o dwóch podstawowych, które są istotne dla naszych dalszych analiz. Są to: (1) charakterystyczne dla tradycji psychologii ekologicznej jest pojęcie afordancji, którym nie posługują się enaktywiści[67]; (2) nacisk psychologów ekologicznych na ustrukturyzowanie i normatywność obecną w środowisku. Wspólne dla obu stanowisk jest, o czym już wspominałem, nastawienie antyreprezentacjonistyczne (por. Chemero, 2014). To ostatnie silnie podkreśla Hutto (2005, s. 389):

> Cechą definiującą niedawny zwrot enaktywistyczny w kognitywistyce jest to, że kwestionuje on paradygmat reprezentacjonistyczny, w zamian podkreślając osadzoną i ucieleśnioną naturę czynności poznawczych. Ruch ten w szczególności wiązał się z odrzuceniem przekonania, według którego możemy zrozumieć podstawy codziennych umiejętności w terminach manipulowania, leżącymi u ich podstaw, reprezentacjami świata (*of a pre-given world*).

[66] Na temat różnic pomiędzy radykalnym a tzw. konserwatywnym enaktywizmem por. Hutto, 2005.
[67] Wyjątkiem jest Bruineberg i współpracownicy (2016), którzy próbują łączyć oba te ujęcia.

Z powodu nastawienia antyreprezentacjonistycznego jestem skłonny traktować podejścia ekologiczne i enaktywistyczne jako odmiany radykalnego odczytania PP, które, przypomnijmy, przeciwstawiam reprezentacjonizmowi ujęcia konserwatywnego. Jako przykład takiego radykalnego ujęcia można wskazać kolejno propozycję Nico Orlandi (2016), która reprezentuje podejście ekologiczne w PP, czy Michaela Kirchhoffa i Iana Robertsona (2018), którzy są przedstawicielami podejścia enaktywistycznego. Propozycja Orlandi jest o tyle interesująca, o ile można na jej przykładzie pokazać możliwość PP, które zrywa z dziedzictwem Helmholtza.

1.5.4. Propozycja Nico Orlandi

Orlandi, wbrew obiegowym opiniom i standardowym wykładniom PP, proponuje ujęcie, które zupełnie zrywa z wywodzącym się z myśli Helmholtza, konstruktywizmem i inferencjalizmem. Autorka broni przekonania, zgodnie z którym PP nie jest inferencjalne i konstruktywistyczne, lecz ekologiczne. Swoje stanowisko uzasadnia na podstawie modeli statystyki obrazów naturalnych (*natural scene statistic*). Statystyki te są rozumiane jako nieprzetworzone obrazy rejestrowane przez aparat zmysłowy człowieka. Charakteryzują się dużą regularnością. Ich definiowanie wymaga uwzględnienia takich czynników jak: rodzaj analizowanej sceny, warunki oświetlenia (tj. kąt padania światła i jego natężenie), pole widzenia, rozdzielczość obrazu. Charakterystyczne dla tych modeli jest ich niezależność od rotacji oraz skali ujęcia, dzięki czemu opis obrazów jest niezależny od ich przekształceń (Lyu, Simoncelli, 2007). Modele te zakładają, że (1) systemy percepcyjne służą do interpretacji obrazów; oraz że (2) wyewoluowały one w reakcji na fizyczne własności środowiska naturalnego (Geisler, Diehl, 2003). Dodatkowo odkrywają w środowisku pewne statystyczne własności, dzięki których regularności możemy postrzegać odpowiednie własności obrazów, a nie zaś, jak utrzymuje konstruktywizm, dzięki ich uprzedniej znajomości. To nie jest tak, że postrzegamy świat jako *tak a tak* ustrukturyzowany ze względu na sposób, w jaki przetwarzamy informacje. To raczej świat, w którym żyjemy, ma już uprzednio pewną określoną regularność, która ma dla nas znaczenie jako ucieleśnionych i osadzonych w środowisku systemów biologicznych (Orlandi, 2016, s. 13).

W tym kontekście trzeba przeanalizować kluczowy dla zrozumienia PP oraz jego ujęć, konstruktywistycznego i ekologicznego, tzw. problem indeterminacji (*inverse problem*) w optyce (por. Hohwy, Roepstorff, Friston, 2008). Dotyczy on sposobu, w jaki obraz jest odbierany przez siatkówkę oka. Mianowicie: określone pobudzenie siatkówki może być spowodowane przez dwa obiekty o różnych rozmiarach, znajdujące się w takiej samej odległości od oka, lub przez dwa obiekty o takich samych rozmiarach, znajdujące się w różnej odległości od oka. Należy

sądzić, że w każdym przypadku możliwy jest tylko jeden scenariusz. Zasadne jest zatem przekonanie, że układ wzrokowy musi „mieć" jakiś sposób na rozwiązanie tego problemu. W ujęciu konstruktywistycznym przyjmuje się, że wykorzystuje on dane z siatkówki oka, interpretując je na podstawie uprzedniej wiedzy, dotyczącej tego, co jest prawdopodobniejsze w środowisku i jakie przyczyny mogły wywołać dane pobudzenie. Wendy J. Adams trafnie to wyraża, mówiąc, że „aby móc zinterpretować złożone i nieokreślone dane wejściowe, ludzki układ wzrokowy wykorzystuje wcześniejszą wiedzę lub założenia dotyczące świata" (Adams, Graf, Ernst, 2004, s. 1057). W takim ujęciu percepcja jest rozumiana inferencyjnie, co już wiadomo także na podstawie wcześniejszych analiz.

Konstruktywistycznej interpretacji problemu indeterminacji w optyce przeciwstawia się interpretację ekologiczną. Zwolennicy teorii ekologicznych odrzucają tezę, według której aparat wzrokowy jest ograniczony do odbierania przez siatkówkę wyłącznie danych statystycznych. Istotny jest bowiem sposób, w jaki światło pada na siatkówkę. To zaś wiąże się ze zmianą układu światła w otoczeniu, która jest powiązana z ruchem. Wraz ze zmianą położenia spostrzegającego niektóre aspekty układu świetlnego zmieniają się, podczas gdy inne pozostają stałe. To właśnie te niezmienne aspekty określają scenę dla spostrzegającego oraz dostarczają wzorce, które można spostrzec lub wyodrębnić z otoczenia na drodze interakcji. Wszelkie informacje konieczne dla spostrzegającego zawarte są w oświetlonym otoczeniu, tj. kształty, wielkość obiektów, głębokość pola widzenia itd. (por. Gibson, 1979). Nie musimy zatem ich wywnioskować, konstruować albo przewidywać.

Modele statystyki obrazów naturalnych są kompatybilne z ujęciem ekologicznym. Z ich perspektywy rozwiązanie problemu indeterminacji polega na sprawdzeniu warunków środowiskowych, które prowadzą do spostrzegania danego przedmiotu lub jego własności. Takie podejście prowadzi do odkrycia, że przykładowo wysokie natężenie światła otacza się częściej tarczami o wysokim natężeniu światła niż tarczami o niskim natężeniu światła. I odwrotnie, otoczenie o niskim natężeniu światła częściej współwystępuje z tarczami o niskim natężeniu światła (Orlandi, 2016, s. 14–15). Percepcja luminancji (tj. miara natężenia oświetlenia padającego w danym kierunku) i jasności światła jest funkcją nie tylko ilości światła odbijanego od powierzchni przedmiotów, ale także własności transmitancji przestrzeni pomiędzy przedmiotami a spostrzegającym (np. ze względu na wpływ atmosfery). Taka sama luminancja może być spowodowana ciemną powierzchnią, która jest jasno oświetlona, albo jasną powierzchnią, słabiej oświetloną. Celem widzenia jest „ustalenie", która kombinacja odbicia światła i oświetlenia odpowiada za określoną ilość luminancji w siatkówce. Podsumowując, z perspektywy prezentowanego tu modelu rozwiązanie problemu indeterminacji opiera się na odwołaniu do tego, jak wygląda świat, który w sposób naturalny ogranicza (*constraints*) potencjalne bodźce wzrokowe oraz hipotezy percepcyjne (Orlandi, 2016, s. 16). Okazuje się, że

nie trzeba przyjmować podejścia konstruktywistycznego, aby wyjaśnić percepcję wzrokową.

Powyższe analizy prowadzą Orlandi do konkluzji, że bayesowski model generatywny w rzeczywistości jest zrelatywizowany do statystyki naturalnych obrazów lub – mówiąc bardziej obrazowo – do środowiska. Dlaczego? Ponieważ z wyjątkiem niektórych wyjściowych hipotez percepcyjnych postuluje się tutaj istnienie struktur, które nie są reprezentacjami, a pewnymi stałymi obecnymi w środowisku. Zatem bayesowskie wnioskowania percepcyjne nie są właściwie żadnymi inferencjami, ale procesami, które są uzbrojone w mechanizmy detekcji. Przez detektory zazwyczaj się rozumie takie stany, jakie monitorują warunki proksymalne, tj. określające, który bodziec czy przedmiot jest bliższy punktowi odniesienia (spostrzegającemu), a także takie, jakie przyczynowo na siebie wpływają. Są one zatem zwykłymi przyczynowymi pośrednikami. Nie modelują żadnych przedmiotów albo relacji pomiędzy nimi oraz działają w bardzo ograniczonym polu. Zachowania systemu, który posługuje się detektorami, opiera się na informacjach pochodzących z monitorowanego otoczenia. To, co dany system może zrobić, wyjaśnia się poprzez odwołanie się do określonej sytuacji środowiskowej, na którą są ustawione detektory[68].

Propozycja Nico Orlandi sugeruje, że można wyjaśnić działanie modelu generatywnego bez powoływania się (1) na inferencjalizm oraz (2) wewnętrzne reprezentacje. Jeżeli tak jest, a wykorzystanie przez nią statystyki obrazów naturalnych tego dowodzi, to okazuje się, że możliwa jest taka wykładnia PP, która zrywa zarówno z konstruktywizmem, jak i reprezentacjonizmem i którą można określić mianem „ekologicznej", o ile z tym terminem wiąże się przekonanie o psychologicznej i epistemologicznej bezpośredniości percepcji oraz przyjęcie tezy antyreprezentacjonistycznej[69].

Analizy Orlandi wykazują, że ani konstruktywizm, ani inferencjalizm, ani reprezentacjonizm nie są *implicite* zawarte w założeniach PP, co mógłby sugerować Zahavi (por. podrozdział *Krytyka przetwarzania predykcyjnego*). Sposób zatem, w jaki zinterpretuje się PP, zależy od szeregu założeń teoretycznych i metateoretycznych. Z tej perspektywy należy stwierdzić, że krytyka Zahaviego oraz innych nie daje się utrzymać i dotyczy w ograniczony sposób tylko wycinka badań nad PP.

[68] Przykładowymi detektorami są niektóre komórki ludzkiego układu odpornościowego.

[69] Teza ta jest kontrowersyjna, ponieważ część badaczy uważa, że można bronić reprezentacjonistycznych teorii ekologicznych (por. Bickhard, Richie, 1983; Norman, 2018).

Tabela 1
Porównanie konserwatywnego i radykalnego PP ze względu na ich inspiracje, teoretyczne zobowiązania i przyjęty model wyjaśniania

	Charakterystyka	Konserwatywne PP	Radykalne PP
Źródła	Helmholtz, nurt psychologii „analiza przez syntezę"	√	X
	Fenomenologia, psychologia ekologiczna, enaktywizm	X	√
Zobowiązania	Reprezentacjonizm	√	X
	Inferencjalizm	√	X
	Internalizm	√	X
Charakter percepcji	Bezpośrednia	X	√
	Pośrednia	√	X
Model wyjaśniania	Mechanistyczny	√	X
	Inne	√	√

Podsumowanie

W niniejszym rozdziale dokonałem przeglądu podstawowych pojęć i stanowisk obecnych w PP. Zrekonstruowałem jego główne źródła i inspiracje. Wykazałem, że rama badawcza PP wykorzystuje szeroko stosowaną w kognitywistyce metodę inżynierii odwrotnej oraz że opiera się na wnioskowaniach abdukcyjnych. Nie jest to zatem koncepcja neokantowska, pomimo wielu jakościowych podobieństw. Następnie zrekonstruowałem zarzut antyrealizmu wobec PP, który sformułował Zahavi. Odpowiadając na jego krytykę, wykazałem, że charakterystyczne dla konserwatywnego PP stanowisko internalistyczne, zakładające, iż granicą systemu poznawczego jest czaszka bądź granica systemu nerwowego, nie implikuje rozstrzygnięć antyrealistycznych, ponieważ model generatywny funkcjonuje w ścisłym związku przyczynowym ze światem w tym sensie, że niektórym wzorcom statystycznym obecnym w sygnale zmysłowym odpowiadają realne struktury w świecie. Wzorce te są reprezentowane przez S-reprezentacje. Na podstawie badań Drayson wykazałem, że trzeba odróżnić trzy znaczenia pośredniości percepcyjnej w PP: psychologiczną, metafizyczną i epistemologiczną. Pośredniość PP zakłada tylko w sensie psychologicznym. Pośredniość psychologiczna percepcji nie warunkuje pośredniości percepcyjnej w sensie metafizycznym bądź epistemologicznym. W dalszej części omówiłem tzw. radykalne wersje PP oraz wskazałem na ich źródła i genezę.

Rozdział II
Przetwarzanie predykcyjne jako bayesowski model eksplanacyjny

Wstęp

W niniejszym rozdziale argumentuję za tym, że PP jest nietrywialną ramą wyjaśniającą poznanie, opartą na algorytmach bayesowskich. Zgodnie z popularną wykładnią PP mózg implementuje probabilistyczną sieć bayesowską, co w praktyce oznacza, że wszelkie procesy poznawcze mają być zastosowaniem reguły Bayesa i dają się opisać w kategoriach rozkładu prawdopodobieństwa. Takie rozwiązanie budzi jednak liczne kontrowersje i wielu autorów argumentuje za tym, że modelowanie bayesowskie ma niską wartość eksplanacyjną i ostatecznie oferuje ono opisy tylko na obliczeniowym poziomie analizy wskazanym przez Marra. Oznacza to, że PP nie ma zastosowania na poziomach algorytmów i implementacji. Twierdzę, że krytyka ta, pomimo iż trafnie charakteryzuje część modeli bayesowskich obecnych w literaturze – jako przykład podaję *analizę racjonalną* – nie jest zasadna w odniesieniu do PP. Po pierwsze dlatego, że istnieje wiele możliwych algorytmów, które pokazują, w jaki sposób to podejście modeluje czynności poznawcze, po drugie zaś istnieją mocne argumenty na rzecz tezy, że ma ono również zastosowanie w odniesieniu do mechanistycznego poziomu implementacji. Argumentuję więc za tym, że PP należy zaliczyć do opisanego przez Matta Jonesa i Bradleya Love'a (2011) bayesizmu oświeconego, który przeciwstawia się słabemu eksplanacyjnie bayesowskiemu fundamentalizmowi[1].

W dalszej części rozważań dyskutuję sformułowany przez Elqayam i Evansa (2011) zarzut, który dotyczy normatywnego charakteru modeli bayesowskich. Badacze ci twierdzą, że należy porzucić myślenie o wyjaśnianiu poznania w kategoriach preskryptywnych i przejść na pozycje deskryptywne. W praktyce oznacza to, że należy odrzucić przekonanie o normatywnym charakterze modeli bayesowskich oraz tezę o racjonalności normatywnej. W dyskusji z tym poglądem wykazuję, że podział na deskryptywizm bayesowski i preskryptywizm bayesowski jest pozorny, co oznacza,

[1] Stanowisko wywodzące się z myśli Bayesa określam jako „bayesizm", a nie jak się również przyjęło pisać w literaturze polskiej – bayesianizm (por. Grobler, 2019; Kawalec, 2000). Mój wybór uzasadniam odwołaniem się do zasad tworzenia przymiotników od nazwisk w języku polskim. Zgodnie z tymi zasadami powinno się pisać „bayesowski", a nie „bayesiański", analogicznie do tego, jak np. pisze się „heglowski", a nie „hegeliański". „Bayesianizm" wydaje się kalką angielskiego „bayesianism". Za „bayesizmem" opowiadają się w Polsce np. Marcin Miłkowski czy Jan Woleński (2001).

że ostatecznie mamy do czynienia z dwoma formami preskryptywizmu, tj. słabą i mocną. Polemika ta pozwala mi zadać pytanie o to, czy powinno się przyjmować realistyczną czy antyrealistyczną postawę wobec bayesowskiego PP. Odpowiedź uzależniam od tego, czy da się wykazać, że mechanizmy poznawcze faktycznie mają charakter bayesowski. Argumentuję za tym, że słaba wersja preskryptywizmu (preskryptywizm epistemiczny) jest zasadna dla PP. Pozwala przyjąć antyrealizm w odniesieniu do PP. Wiąże się bowiem z możliwością wyjaśniania zjawisk przy użyciu twierdzenia Bayesa, nie przesądza jednak, czy zjawiska te są bayesowskie *per se*. Pełne uzasadnienie realizmu w odniesieniu do bayesowskiego PP zakłada jednak przyjęcie mocnego preskryptywizmu, który określam jako „ontyczny". Ontyczny preskryptywizm zakłada, że zjawiska wyjaśnia się za pomocą twierdzenia Bayesa, ponieważ są one bayesowskie. Jeżeli są one bayesowskie z natury, to znaczy, że *powinno* się je wyjaśniać przy użyciu modelowania bayesowskiego. Temu zagadnieniu poświęcam analizy w rozdziałach *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego* i *Mechanizmy normatywne a działania w przetwarzaniu predykcyjnym*.

2.1. Uwagi wprowadzające

Przypomnijmy: wedle hipotezy proponowanej przez zwolenników koncepcji PP (Clark, 2013, 2016; Friston, 2010, 2012; Hohwy, 2013, 2014; Metzinger, Wiese, 2017) podstawowym zadaniem mózgu jest minimalizacja błędów predykcyjnych. Mózg w sposób ciągły tworzy statystyczne przewidywania, które są hierarchicznie zorganizowane i tworzone na poszczególnych poziomach modelu generatywnego. Następnie za ich pomocą zostaje przewidziany obecny i przyszły kształt danych docierających za pośrednictwem wejść zmysłowych. Hierarchiczny model koresponduje z hierarchią zmiennych losowych odbieranych przez zmysły (Metzinger, Wiese, 2017, s. 3). Predykcje są zatem z jednej strony warunkowane posiadanym przez dany system poznawczy wewnętrznym modelem świata, z drugiej zaś przez zmienne informacje napływające ze świata. Obie te warstwy wzajemnie się ograniczają, pozostając w dynamicznych zależnościach (Hohwy, 2013, s. 69–70). Opierając się na wiedzy dotyczącej przyczynowych zależności w świecie, model generatywny stawia hipotezy odnoszące się do prawdopodobieństwa zajścia pewnych zdarzeń. Dzięki temu może zminimalizować błędy predykcyjne[2].

Ustalenia te biorą się z obserwacji, że informacje zmysłowe nie kształtują percepcji w sposób bezpośredni. Są one raczej w sposób aktywny dobierane i odpowiednio

[2] Według hipotezy predykcyjnej mózg generuje odpowiednie predykcje, szacując poziom swojej niepewności (Clark, 2016, s. 57). Szacowanie niepewności wiąże się wprost z tzw. precyzją. Im większą precyzję system przypisuje danemu błędowi predykcyjnemu, tym mniejszy jest ten błąd (Friston, 2010).

wykorzystywane. Predykcje kierują tym, co jest spostrzegane, ale także sposobem integracji odpowiednich aspektów spostrzeganego świata. Problem percepcji dotyczy zatem możliwości wykorzystania danych, które dochodzą do mózgu za pośrednictwem wejść zmysłowych w taki sposób, aby organizm nie popełniał błędów predykcyjnych. Błędy te zaś biorą się z nieznajomości przyczynowych źródeł pobudzeń zmysłowych. Nie istnieje jedno-jednoznaczna relacja zachodząca pomiędzy przyczynami i skutkami – różne przyczyny mogą wywoływać takie same skutki. Nieznajomość tych przyczyn w praktyce może stanowić zagrożenie dla danego organizmu. Z tego powodu Hohwy powiada, że koncepcja mózgu minimalizującego błąd predykcyjny nie wyjaśnia faktu percepcji, ale sposób, w jaki ona się odbywa.

Dla wytłumaczenia natury percepcji kluczowa okazuje się kwestia predykcji generowanych przez system poznawczy podczas każdego kontaktu z rzeczywistością. W tej perspektywie problem percepcji ogniskuje się wokół pytania o to, w jaki sposób kształtowane są i wybierane prawdziwe hipotezy dotyczące świata (Hohwy, 2013, s. 16)? Inaczej mówiąc: w jaki sposób, nie znając przyczyn pobudzeń zmysłowych, mózg może na ich bazie wyprowadzać odpowiednie hipotezy? Zadanie takie powinno polegać na ekstrapolacji określonych informacji z niepewnego zbioru danych. Przetwarzanie informacji w mózgu opiera się zatem na jakiejś formie wnioskowania statystycznego.

Część badaczy (m.in. Harkness, Keshava, 2017; Hohwy, 2013, 2016) przyjmuje hipotezę bayesowskiego mózgu, wedle której ośrodkowy układ nerwowy konstruuje i testuje wewnętrzne modele świata zewnętrznego na drodze realizacji procesów poznawczych, stanowiących aproksymację[3] wnioskowania bayesowskiego (Clark, 2013b, s. 189). Następnie, wykorzystując hierarchicznie zorganizowane inferencje, mózg tworzy odpowiednie wielopoziomowe modele generatywne, które odgórnie generują hipotezy „interpretujące" oddolnie przychodzące informacje pochodzące z sygnału zmysłowego. Każdy poziom takiego modelu minimalizuje błędy predykcyjne na poziomie niższym: od procesów neuronalnych, aż po wyższe procesy poznawcze.

Co to znaczy? Rozważmy pewien przykład: jadę nocą nieoświetloną drogą jednokierunkową. Dostrzegam dwa zbliżające się punkty świetlne. Przewiduję, że są to światła samochodu, który jedzie z naprzeciwka. Zakładam również, że ten samochód jedzie właściwym pasem. Istnieje także ryzyko, że jedzie on pod prąd. Nie mam jednak pewności. W jaki zatem sposób mogę podjąć decyzję dotyczącą tego, co mam zrobić? Jeżeli moja reprezentacja ruchu wykonywanego przez samochód jadący naprzeciw mnie okaże się błędna, to błąd ten ostatecznie może mnie kosztować życie. Trafność utworzonej przeze mnie predykcji dotyczącej tego, co się zaraz wydarzy na drodze, nie tylko zależy od informacji, które do mnie docierają za pośrednictwem zmysłów, ale także mojej wiedzy, doświadczenia i przewidywań, których dokonuję.

[3] Aproksymacja jest tutaj rozumiana jako przybliżanie danej funkcji przy użyciu innej, prostszej, która jest łatwiejsza do zbadania i zastosowania.

Błąd predykcyjny zostanie zminimalizowany dopiero wtedy, gdy model przyjmie najlepszą możliwą hipotezę dotyczącą przyczyn źródła sygnału zmysłowego. Na podstawie tej hipotezy tworzone są następnie predykcje warunkujące działania danego systemu poznawczego czy organizmu. Model, co podkreślałem, jest wielopoziomowy, dzięki czemu każdy jego poziom minimalizuje błąd predykcyjny na poziomie niższym. W przykładzie z samochodem jeden poziom modelu (wyższy) dotyczyć będzie możliwości rozpoznania świetlnych punktów jako świateł samochodowych, inny (niżej leżący) będzie odnosił się np. do detekcji krawędzi spostrzeganego przedmiotu, kolejny poziom będzie generował predykcje dotyczące np. czasu, w jakim może dojść do kolizji obu pojazdów. Na każdym poziomie model szacuje, na ile precyzyjny jest dany błąd predykcyjny, tak aby można było dokonać rewizji dotychczas przyjętych hipotez (por. Friston, 2009, s. 299). Jak zatem możliwy jest proces tworzenia i selekcji hipotez, które mają różne wartości prawdopodobieństwa? Większe prawdopodobieństwo ma hipoteza głosząca, że światła z naprzeciwka wywołane są przez inny samochód, niż hipoteza, według której z naprzeciwka nadlatuje statek kosmiczny.

Postulowany przez wielu badaczy pogląd, że mózg predykcyjny jest mózgiem bayesowskim, oznacza, że model generatywny przypomina w jakimś stopniu zaimplementowaną w mózgu hierarchiczną sieć bayesowską[4]. Ogólna idea jest taka: model generatywny tworzy odpowiednie hipotezy i dokonuje ich wyboru, stosując przy tym wnioskowanie Bayesa. Zatem minimalizacja błędów predykcyjnych jest empirycznym zastosowaniem reguły Bayesa. Ujęcie takie wiąże się bezpośrednio z przekonaniem Helmholtza, który głosi, że percepcja jest nieświadomym wnioskowaniem statystycznym (Helmholtz, 1867; Hohwy, 2013, rozdział 1; Rescorla, 2015, s. 1 i nn.). Wielu badaczy jest zgodnych co do tego, że procesy poznawcze opierają się na jakiejś formie wnioskowania probabilistycznego. Pytanie zatem brzmi: dlaczego to właśnie wnioskowanie bayesowskie ma być tym, które wykorzystuje system poznawczy? Aby na nie odpowiedzieć, należy przejść do szczegółowych rozważań. Na wstępie trzeba jednak wprowadzić podstawowe pojęcia i rozróżnienia związane z teorią prawdopodobieństwa, istnieje bowiem kilka możliwych interpretacji tych pojęć, jak i samego bayesizmu.

2.2. Podstawowe pojęcia

Kluczowe dla analiz prezentowanych w niniejszej pracy jest pojęcie prawdopodobieństwa. Jest to pojęcie wykorzystywane w wielu dziedzinach nauki. Potocznie jego znaczenie wiąże się z szansą na to, że zajdzie jakieś zdarzenie. Takie rozumienie odnosi się do oczekiwania, że coś się wydarzy. W sensie statystycznym

[4] Określa się je również mianem „sieci przekonań" (*belief networks*). Por. Pearl, 1988, s. 104 i nn.

prawdopodobieństwo wiąże się z miarą lub miarami służącymi do opisu pewności bądź częstości (np. zachodzenia) danego zdarzenia. W tym sensie nie mówi się jednak o prawdopodobieństwie określonych, indywidualnych zdarzeń, a pewnych ich typów bądź rodzajów (np. takich, jakie polegają na wyrzuceniu reszki monetą). Typ, *resp*. rodzaj, zdarzeń interpretuje się jako zbiór zdarzeń określonego rodzaju (np. zbiór wyników odnoszących się do wyrzucenia reszki monetą). W praktyce oznacza to, że jeżeli zastosuje się do takiego zbioru zdarzeń rachunek prawdopodobieństwa, to zostanie wyróżniony określony zbiór zdarzeń (tzw. elementarny), którego elementami są zdarzenia indywidualne (np. jednostkowe wyrzucenie reszki przez Jana). Zdarzenia, które są pewne, traktuje się jako zbiór pełny (w odniesieniu do zbioru wszelkich zdarzeń elementarnych), zaś zdarzenia niemożliwe składają się na zbiór pusty (Mortimer, 1982, s. 22).

Tak rozumiane prawdopodobieństwo definiuje się jako funkcję o wartościach liczbowych, której zbiór argumentów, tj. przedmiotów, dla jakich jest określone to prawdopodobieństwo, tworzy podzbiór Boole'a, spełniający twierdzenia rachunku prawdopodobieństwa. Tak rozumianą funkcję oznacza się symbolem P. I tak P(A) oznacza prawdopodobieństwo A. Tak rozumiane prawdopodobieństwo ma określoną wartość liczbową. Istnieje wiele definicji prawdopodobieństwa, jednak w niniejszej pracy odnoszę się do pojęcia prawdopodobieństwa, które jest definiowane przez aksjomaty Andrieja Kołmogorowa. Aksjomaty te nakładają na prawdopodobieństwo następujące warunki:

- dla dowolnego elementu x: $0 \leq P(A) \leq 1$ (aksjomat nienegatywności);
- $P(T) = 1$, gdzie T to tautologia (aksjomat normalizacji);
- $P(A \vee B) = P(A) + P(B)$, gdy $A \wedge B = \neg T$ (aksjomat skończonej addytywności).

Aksjomaty 1 i 2 określają, że prawdopodobieństwo dowolnego zdarzenia jest liczbą z przedziału od zera do jedności (przy czym zdarzenie pewne ma prawdopodobieństwo maksymalne równe jedności, zaś zdarzenie niemożliwe ma prawdopodobieństwo równe zeru). Aksjomat 3 stwierdza, że prawdopodobieństwo zajścia jednego ze zdarzeń parami rozłącznych (tj. takich zbiorów, jakie nie mają wspólnych elementów) jest równe sumie prawdopodobieństwa poszczególnych zdarzeń.

Powyższe aksjomaty dotyczą tzw. prawdopodobieństwa pierwotnego (absolutnego), tj. prawdopodobieństwa *a priori*: P(A). Za jego pomocą definiuje się prawdopodobieństwo warunkowe (*a posteriori*), które jest już funkcją dwuargumentową: P(A/B). Wyrażenie to oznacza wartość prawdopodobieństwa warunkowego dla argumentu A ze względu na argument B lub innymi słowy, prawdopodobieństwo zajścia A, ze względu na prawdopodobieństwo zajścia B. Prawdopodobieństwo warunkowe definiuje się przez prawdopodobieństwo pierwotne:

$$\mathbf{P}(\boldsymbol{A}|\boldsymbol{B}) = \frac{\mathbf{P}(\boldsymbol{A} \wedge \boldsymbol{B})}{\mathbf{P}(\boldsymbol{B})}$$

Należy dodać, że P(*B*) nie jest równe zero. Jeżeli zaś jest, to wartość P(*A*/*B*) nie jest możliwa do określenia.

Na podstawie aksjomatów Kołmogorowa, definicji prawdopodobieństwa absolutnego i warunkowego dowodzi się twierdzenia Bayesa[5]:

$$\mathbf{P}(\boldsymbol{A}|\boldsymbol{B}) = \frac{\mathbf{P}(\boldsymbol{B}|\boldsymbol{A})\mathbf{P}(\boldsymbol{A})}{\mathbf{P}(\boldsymbol{B})}$$

gdzie:

A – hipoteza;

B – obserwacja (np. pewnego zbioru danych);

$P(A|B)$ – prawdopodobieństwo wtórne zajścia zdarzenia A przy zajściu warunku B;

$P(A)$ – prawdopodobieństwo pierwotne zdarzenia A;

$P(B|A)$ – prawdopodobieństwo wtórne obserwacji względem hipotezy (*likelihood*) – prawdopodobieństwo warunkowe B pod warunkiem prawdziwości zdarzenia A;

$P(B)$ – prawdopodobieństwo pierwotne obserwacji (*marginal likelihood*);

$\frac{P(B|A)}{P(B)}$ – prawdopodobieństwo warunkowe, tj. stopień potwierdzenia hipotezy przez obserwacje.

Twierdzenie Bayesa (zwane również „regułą Bayesa") umożliwia obliczenie prawdopodobieństwa wtórnego P(*A*/*B*), np. danej hipotezy, gdy znane są prawdopodobieństwo odwrotne P(*B*/*A*) oraz prawdopodobieństwa pierwotne P(*A*) i P(*B*). Innymi słowy, twierdzenie Bayesa określa, jaki wpływ na prawdopodobieństwo hipotezy *A* mają dowody bądź dane (np. empiryczne).

Prawdopodobieństwo wtórne P(*A*|*B*) oblicza się, wykonując kolejne kroki reguły (Tyszka, Zaleśkiewicz, 2001, s. 207–208):

1. Określa się prawdopodobieństwo pierwotne danej hipotezy P(*A*).
2. Zbiera się dane *B* i określa (1) ich prawdopodobieństwo pierwotne P(*B*) oraz (2) prawdopodobieństwo wtórne zebranych danych P(*B*/*A*) pod warunkiem prawdziwości danej hipotezy *A*.
3. Oblicza się prawdopodobieństwo warunkowe P(*A*|*B*), które wyraża warunkowe prawdopodobieństwo hipotezy *A* w świetle danych *B* (zgodnie z zapisem reguły Bayesa).

Należy stwierdzić, że bayesizm (rozumiany jako bayesowska teoria konfirmacji) jest ilościową i normatywną teorią racjonalności naukowej analizowaną w aspekcie

[5] Omawiamy przykład odnosi się do obliczenia prawdopodobieństwa prawdziwości danej hipotezy ze względu na zebrane dane.

synchronicznym oraz diachronicznym. Aksjomatycznie zdefiniowane pojęcie prawdopodobieństwa oraz wyprowadzone z niej twierdzenie Bayesa określają warunki tzw. racjonalności synchronicznej dla twierdzeń bądź hipotez naukowych. Zmiana zaś tych twierdzeń, *resp.* hipotez, wiąże się z tzw. racjonalnością diachroniczną. Zwolennicy bayesizmu sposób aktualizacji prawdopodobieństwa, *resp.* zmiany określonych hipotez pod wpływem nowych danych, wiążą z tzw. zasadą warunkowania. Samo określanie prawdopodobieństwa danej hipotezy (aspekt synchroniczny) nie budzi większych wątpliwości (zakładając wiele możliwych interpretacji pojęcia prawdopodobieństwa, o czym wspomnę niebawem), tak możliwość ich aktualizowania rodzi wiele kontrowersji (por. Kawalec, 2003). Jest to istotne dla analiz przeprowadzanych tutaj, ponieważ zgodnie z bazowymi założeniami PP model generatywny nie tylko oblicza prawdopodobieństwo określonych hipotez i przekonań, ale także ma aktualizować swoje parametry na drodze zmiany lub modyfikacji wcześniej uzyskanej wiedzy. Aby móc wskazać, jak to jest możliwe, należy powiedzieć kilka słów, jak w ramie PP jako jednej z możliwych interpretacji bayesizmu jest rozumiane prawdopodobieństwo.

Rachunek prawdopodobieństwa ma różne zastosowanie i zależnie od tego są różnie interpretowane wartości prawdopodobieństwa, tj. stopnie nasilenia pewnych własności stopniowalnych, które przysługują przedmiotom będącym argumentami tej funkcji (Mortimer, 1982, s. 29). Wyróżnia się różne interpretacje pojęcia prawdopodobieństwa w zależności od tego, jaką własność ma mierzyć prawdopodobieństwo:

1. Częstościowa (m.in. Ludwig von Mises): odnosi się do pojęcia częstości względnej danego zdarzenia w pewnym skończonym zbiorze zdarzeń. Jeżeli pewien zbiór X ma n elementów, wśród których jest m elementów mających własność Y, to częstość względna Y w zbiorze X wynosi m/n. Utożsamia się tutaj prawdopodobieństwo danego zdarzenia z granicą jego częstości względnej w pewnym skończonym zbiorze zdarzeń.

2. Logiczna (obiektywna) (m.in. John M. Keynes, Rudolf Carnap; Jaako Hintikka): twierdzi się tutaj, że wartość prawdopodobieństwa zależy od określonych związków logicznych pomiędzy zdaniami (np. wynikanie logiczne). Z tego też powodu, aby określić, że jakieś zdanie wynika z innego, nie jest potrzebna żadna (empiryczna) wiedza na temat przedmiotów, o których są te zdania. Zwolennicy tego ujęcia rezygnują z prawdopodobieństwa częstościowego ze względu na jego empiryczny i indukcyjny charakter oraz twierdzą, że należy znaleźć taką interpretację prawdopodobieństwa, jaka umożliwia logiczną metodę wyznaczania jego wartości[6].

3. Epistemologiczna (m.in. Kazimierz Ajdukiewicz, Henry E. Kyburg): proponuje się, aby wartości prawdopodobieństwa logicznego utożsamić z wartościami prawdopodobieństwa statystycznego, w ten sposób nadając temu pierwszemu empiryczny charakter. Wartości prawdopodobieństwa są zatem szacowane na

[6] Trudności tego stanowiska szczegółowo omawia Halina Mortimer (1982, s. 63–97).

podstawie obserwowanych częstości, czyli indukcyjnie. Przykładowo Kyburg zaproponował, aby prawdopodobieństwo logiczne potraktować nie jako relację pomiędzy zdaniami, ale relację pomiędzy zdaniem a aktualną wiedzą. Oznacza to, że wartość prawdopodobieństwa logicznego jest zależna od wiedzy o prawdopodobieństwach statystycznych, ale nie jest utożsamiona z ich wartościami. Takie ujęcie ma gwarantować niearbitralny charakter wartości prawdopodobieństwa.

4. Subiektywna (Frank Ramsey; Bruno de Finetti; Leonard Savage oraz zwolennicy PP): prawdopodobieństwo jest tutaj rozumiane jako miara stopnia przekonania danej osoby o prawdziwości określonej hipotezy, zdania czy teorii. Samo to stanowisko nie jest oczywiście jednorodne i można wydzielić jego różne podtypy, np. behawioralny, który wiąże przekonania z określonymi zachowaniami w sytuacjach podejmowania decyzji.

W PP przekonania rozumie się na wiele sposobów (por. Sims, 2017), przy czym na potrzeby niniejszych rozważań będę przez nie rozumiał określone hierarchicznie zorganizowane rozkłady prawdopodobieństwa dotyczące nieznanych stanów świata. Z tego też powodu regułę Bayesa będę określał jako „regułę prawdopodobieństwa subiektywnego”[7].

Przykład z nieoświetloną drogą jednokierunkową dobrze ilustruje omawiany tu problem. Przypomnijmy: kiedy jadę autem, spostrzegam z naprzeciwka zbliżające się do mnie światła. Przyjmuję hipotezę, że są to światła samochodu jadącego pod prąd. Zakładam, że jeżeli w nocy widzę światła zbliżające się w moim kierunku (jest to informacja pochodząca z sygnału zmysłowego), to są to światła samochodu jadącego pod prąd (hipoteza wyjaśniająca źródło danej informacji). Taka hipoteza w świetle mojej wcześniejszej wiedzy i doświadczenia jest prawdopodobniejsza niż hipoteza głosząca, że źródłem zbliżających się świateł jest statek kosmiczny albo instalacja świetlna.

Każdy z poziomów modelu generatywnego, który jest zaangażowany w sytuację percepcyjną opisaną powyżej, stosuje regułę Bayesa. Informacja o błędzie modulowana przez mechanizm uwagi (oczekiwana precyzja) trafia do wyższego poziomu, gdzie następuje obliczenie prawdopodobieństwa wtórnego, zmiana parametru modelu oraz utworzenie odpowiedniej predykcji, która zostaje przetestowana. Można tutaj powiedzieć, że zmieniają się parametry wewnętrznej mapy probabilistycznej środowiska. Mózg, opierając się na dostępnych danych odniesionych do omawianej sytuacji, oblicza prawdopodobieństwo predykcji dla określonych poziomów modelu. Dotyczą one m.in.:

- hipotez odnoszących się do kształtów, kolorów czy odgłosów (niskie poziomy modelu);

[7] Istnieje wiele przykładów, które sugerują, że ludzie nie stosują reguły Bayesa, intuicyjnie oceniając prawdopodobieństwo danych zdarzeń w świetle nowych doświadczeń. Popełniają bowiem dwa błędy: (1) są konserwatywni w rewidowaniu znanych prawdopodobieństw pod wpływem nowych danych; oraz (2) ignorują prawdopodobieństwa pierwotne (Tyszka, 2010, s. 232).

- hipotez odnoszących się do ruchu samochodów, długości hamowania, możliwości wymijania itd.;
- hipotez odnoszących się do sytuacji na drodze, poruszających się pojazdów, sygnalizacji świetlnej itd.

Generowane hipotezy opierają się na odbieranych informacjach pochodzących ze zmysłów zewnętrznych oraz zmysłu kinestetycznego[8]. Ważna jest również wiedza zakodowana w pojęciach oraz samowiedza dotycząca podmiotu działania i spostrzegania.

Istotne w naszym przykładzie jest to, że prawdopodobieństwo hipotezy, która ma możliwie najlepiej zminimalizować błąd predykcyjny, opiera się na subiektywnej ocenie danego zdarzenia, a nie np. na odniesieniu do jakichś „obiektywnych" własności rzeczywistości (Rescorla, 2015, s. 4). Owa ocena nie jest jednak zupełnie arbitralna, bowiem wiąże się z probabilistycznym charakterem informacji zawartej w sygnale zmysłowym. Jest jednak subiektywna, ponieważ, o czym już wspominałem, nie są nam znane przyczynowe źródła wywołujące pobudzenia zmysłowe. Należy zatem przyjąć ich najprawdopodobniejsze (z punktu widzenia organizmu) wyjaśnienie.

W niniejszej pracy regułę Bayesa traktujemy jako dotyczącą optymalnego czy racjonalnego wyboru działań (por. Orlandi, 2017, s. 8). Mówiąc zatem o przekonaniach, rozpatrujemy je z punktu widzenia możliwości podejmowania odpowiednich działań. W tym sensie wnioskowanie bayesowskie jest metodą racjonalnego wiązania istniejących informacji, z których przynajmniej część jest niepewna, z nowymi danymi. Niepewność informacji oznacza, że można ją opisać przy użyciu rozkładu prawdopodobieństwa.

Wraca w tym miejscu anonsowany już wcześniej przeze mnie problem zasady warunkowania, która umożliwia aktualizowanie prawdopodobieństwa określonych hipotez bądź przekonań. Zasada ta polega na systematycznym zastosowywaniu reguły Bayesa w taki sposób, że prawdopodobieństwa względne hipotez z wcześniejszych etapów uzasadniania są traktowane jako prawdopodobieństwa pierwotne w następnych etapach:

$$\mathrm{P}(A/B) = \frac{\mathrm{P}(A)\mathrm{P}(B|A)}{\mathrm{P}(B)}, \text{ niech } \mathrm{P}(A/B) = \mathrm{P}(A1) \Rightarrow$$

$$\Rightarrow \mathrm{P}(A1/B1) = \frac{\mathrm{P}(A1)\mathrm{P}(B1|A1)}{\mathrm{P}(B1)} \text{ itd.}$$

Teoria konfirmacji oparta na tak definiowanej zasadzie warunkowania służy zwolennikom bayesizmu (szczególnie w wersji subiektywnej) do wykazywania obiektywności naukowej. Niezależnie bowiem od preferencji i przekonań naukowców w stosunku do określonej hipotezy (tzn. prawdopodobieństwo pierwotne, jakie

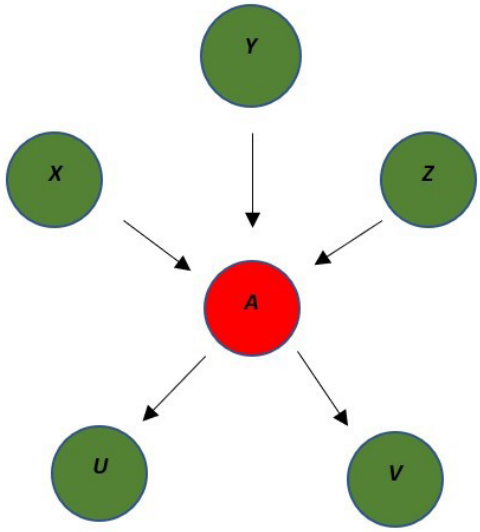
[8] Tak zwana priopercepcja – dane dotyczące stanów ciała, ułożenia kończyn, lokalizacji ciała itd.

przypisują jej naukowcy), przy uwzględnieniu tych samych danych (np. empirycznych), po jakimś czasie różnice w preferencjach naukowców zostaną zniwelowane (tzn. prawdopodobieństwa wtórne będą niemal identyczne) (Kawalec, 2003, s. 116)[9].

Zanim przejdziemy dalej, konieczne jest jeszcze wprowadzenie formalnego pojęcia sieci bayesowskiej i koców Markowa, które są konieczne dla zrozumienia wielu rozstrzygnięć w obrębie debaty nad PP.

Sieć bayesowska wykorzystywana jest do przedstawiania zależności pomiędzy zdarzeniami, zgodnie z aksjomatami rachunku prawdopodobieństwa. Siecią bayesowską (Jensen, 1996) nazywamy parę (*D*, *P*), gdzie *D* jest skierowanym grafem acyklicznym (*directed acyclic graph* – DAG), a *P* rozkładem prawdopodobieństwa. Każdy wierzchołek w sieci przechowuje rozkład $P(Xi|\pi(i))$, gdzie $X\pi(i)$ jest zbiorem wierzchołków odpowiadających $\pi(i)$ – poprzednikom (rodzicom) wierzchołka (i)[10]. W takim grafie:

- węzły reprezentują zmienne losowe (własność danego przedmiotu);
- łuki reprezentują określone (probabilistyczne) zależności pomiędzy zmiennymi (zmienna *x* ma taki a taki wpływ na zmienną *y*);
- zmienne reprezentowane przez określone węzły przyjmują własności dyskretne (np. TAK, NIE);
- każdy określony węzeł ma powiązaną z nim tablicę prawdopodobieństw warunkowych określających wpływ wywierany na ten węzeł przez jego poprzedników (rodziców) w grafie.



Rysunek 3. Acykliczny graf skierowany sieci bayesowskiej

[9] Problemy teorii konfirmacji omawiam w podrozdziale *Problemy z modelami bayesowskimi*.

[10] W grafie wierzchołki są etykietowane nazwami atrybutów. Przy każdym wierzchołku występuje tabela prawdopodobieństw warunkowych pomiędzy danym wierzchołkiem a jego rodzicami.

W PP zmienne *X*, *Y*, *Z* reprezentują określone dane (np. nowe informacje pojawiające się w sygnale zmysłowym) oraz nabyte przekonania (*priors*), zmienna *A* nową predykcję, zaś zmienne *U*, *V* nowe hipotezy (przekonania). Należy tutaj stwierdzić, że celem propagacji niepewności w sieciach bayesowskich jest obliczenie prawdopodobieństwa warunkowego pewnej zmiennej (zwykle oznaczającej pewną hipotezę) względem zadanych wartości innych zmiennych (np. obserwacji, nowych danych). W takim grafie węzeł *X* jest rodzicem lub poprzednikiem wierzchołka *A*, a wierzchołek *A* jest potomkiem lub następnikiem węzła *X*, jeżeli istnieje bezpośrednia krawędź z wierzchołka *X* do *A*. Oznacza to, że prawdopodobieństwo pojawienia się wierzchołka potomnego zależy tylko od jego rodziców.

Trzeba pamiętać o tym, że aby móc skonstruować jakąkolwiek sieć bayesowską, należy podjąć kilka kroków: zdefiniować zmienne; zdefiniować połączenia pomiędzy tymi zmiennymi; określić prawdopodobieństwa pierwotne, wtórne i warunkowe; wprowadzić odpowiednie dane do sieci; określić sposób ich aktualizacji[11]. Warto przypomnieć, że zgodnie z podstawową wykładnią PP model generatywny, *resp.* mózg, implementuje probabilistyczną sieć bayesowską, co w praktyce oznacza, że wszelkie procesy poznawcze mają być zastosowaniem reguły Bayesa i dają się opisać w kategoriach rozkładu prawdopodobieństwa.

Znając już pojęcie sieci bayesowskiej, mogę wprowadzić pojęcie koców Markowa, które pojawia się w wielu dyskusjach dotyczących szeregu zagadnień i problemów w ramach PP i nie tylko (por. podrozdział *Internalizm a koce Markowa* oraz podrozdział *Koce Markowa a przetwarzanie predykcyjne*).

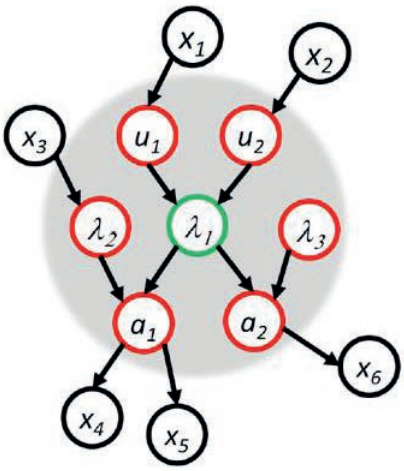
Pojęcie łańcuchów Markowa wywodzi się z prac Andrieja Markowa, badającego abstrakcyjne systemy, które pamiętają swoje przeszłe trajektorie tylko wtedy, gdy przechowują bieżące sygnały. Łańcuchem Markowa pierwszego rzędu ($n = 1$) jest taki system, w jakim prawdopodobieństwo każdego zdarzenia zależy wyłącznie od wyniku zdarzenia poprzedniego[12]. Łańcuchy wiążą się z istnieniem procesów stochastycznych, czyli takich, jakie odnoszą się do danej rodziny zmiennych losowych, dających się określić na pewnej przestrzeni probabilistycznej. Przykładem procesu stochastycznego jest rzut monetą. Procesy stochastyczne, które spełniają własność Markowa, nazywa się „procesami Markowa". Przyjrzyjmy się bliżej tym zagadnieniom.

Zwolennicy PP nie mówią wprost o łańcuchach Markowa. Posługują się raczej wprowadzonym przez Judeę Pearl w pracy *Probabilistic reasoning in intelligent systems: Networks of plausible inference* pojęciem koców Markowa (*Markow blankets*) (por. Pearl, 1988, s. 120–121; zob. także podrozdział *Internalizm a koce*

[11] Nie jest jednak proste. Trudności obliczeniowe związane z takim rozumieniem sieci bayesowskiej w PP omawiają Kwisthout i van Rooij (2019).

[12] W ogólności łańcuch Markowa definiuje zależność tylko od *n* (dla ustalonego *n*) poprzednich kroków czasowych.

Markowa)[13]. W sieciach bayesowskich koce Markowa opisuje się, używając pojęć rodziców i dzieci. Wyjaśnienie zachowania danego dziecka wymaga odniesienia się do zachowania jego rodzica i innych dzieci. Nie potrzeba zatem znać stanów węzłów poprzedzających rodziców danego dziecka, tj. dziadków, pradziadków itd. W praktyce oznacza to, że jeżeli chce się przewidzieć stan danego węzła, wszystko, co trzeba znać, to stany węzłów, które tworzą jego koc Markowa. Poniższa ilustracja dobrze to obrazuje:



Rysunek 4. Acykliczny graf skierowany koca Markowa (Hohwy, 2017, s. 3)

Zachowanie węzła λ_1 (rodzic) jest przewidywalne na podstawie znajomości zachowania węzłów z nim sąsiadujących: λ_2, λ_3 (rodzeństwo λ_1); u_1, u_2 (rodzice λ_1) oraz a_1, a_2 (dzieci λ_1). Nie trzeba nic wiedzieć o zachowaniach węzłów położonych na zewnątrz koca (węzły oznaczone jako x_x). Innymi słowy, znajomość stanów węzłów zewnętrznych nie jest konieczna dla znajomości stanów węzłów wewnętrznych. Koc Markowa (k) wokół stanów wewnętrznych λ_n – gdzie wszystkie inne (zewnętrzne) zmienne są oznaczone przez x_n – jest zdefiniowany jako zbiór zmiennych, które uniezależniają λ warunkowo od stanów zewnętrznych. Matematycznie jest to zapisywane w następujący sposób:

$$\lambda \perp x|k \Leftrightarrow P(\lambda, x|k) = P(\lambda|k)\ P(x|k)$$

[13] Systemy biologiczne i sztuczne można satysfakcjonująco modelować jako sieci składające się z węzłów (*node*), które reprezentują składowe jednostki systemu, oraz krawędzi (*edge*), które oznaczają interakcje pomiędzy węzłami. W odniesieniu do mózgu węzły mogą oznaczać neurony, ich populacje albo całe obszary mózgu w zależności od skal przestrzennych, którymi są zainteresowani badacze. Krawędzie zaś mogą reprezentować strukturalne lub funkcjonalne relacje wiążące poszczególne węzły (Rubinov, Sporns, 2010; Sporns, 2014).

Równanie to ilustruje strukturę zależności w rozkładzie połączeń uwarunkowanych stanami całkowitymi na dwa niezależne rozkłady. Z definicji dwie zmienne są warunkowo niezależne wtedy i tylko wtedy, gdy ich wspólne prawdopodobieństwo, uwarunkowane jakąś trzecią zmienną, jest równe iloczynowi ich indywidualnego prawdopodobieństwa uwarunkowanego tą trzecią zmienną. Często mówi się o zmiennych losowych oddzielonych w ten sposób kocami Markowa i związanych z nimi zależnościach warunkowych (w kategoriach rodziców i ich dzieci) (Hipólito i in., 2020, s. 4).

Tak rozumiane koce Markowa służą do wyznaczania statystycznej granicy, która pośredniczy w interakcjach pomiędzy systemem (jego wnętrzem) a środowiskiem (tj. tym, co na zewnątrz systemu) (por. podrozdział *Internalizm a koce Markowa*). Pomysł ten ma zastosowanie m.in. przy opisie i analizie systemów neuronalnych, tj. architektury mózgu w różnych skalach, ale także pozwala na modelowanie struktury komórek, żywych organizmów, sieci społecznych itd. (por. podrozdział *Koce Markowa a przetwarzanie predykcyjne*).

Po tych wstępnych ustaleniach terminologicznych możemy przejść dalej, do szczegółowszej charakterystyki bayesowskiego PP.

2.3. Teza mózgu bayesowskiego

Sascha B. Fink i Carlos Zednik (2017) zauważają, że istnieją dwa oddzielne metodologicznie sposoby modelowania mózgu, pozwalające mówić o nim jako o bayesowskim[14]. Podkreślają przy tym, że sposoby te należy raczej rozważać jako specyficzne ramy (*framework*) wyznaczające badania niż konkretne modele posiadające moce eksplanacyjne. Owe ramy niejako wprowadzają narzędzia do badania odpowiednich zjawisk oraz określają dobór metod adekwatnych do tworzenia takich, a nie innych modeli. Modele zaś w przeciwieństwie do ram wiążą i ustawiają dostępne parametry w celu wyjaśnienia określonego zjawiska.

Pierwszą z ram wskazanych przez Finka i Zednika jest (bayesowska) *analiza racjonalna*. Przy jej pomocy można scharakteryzować i wyjaśnić zachowania oraz zjawiska poznawcze jako formy optymalnego wnioskowania probabilistycznego. W tym ujęciu zachowanie danego organizmu jest optymalnym rozwiązaniem określonego wnioskowania probabilistycznego, które dotyczy zadania wynikającego z kontaktu ze środowiskiem. Tworzony jest tutaj formalny model środowiska, w którym dane zadania są rozwiązywane w możliwie optymalny sposób. Jest to

[14] Można, jak sądzę, wskazać jeszcze inne sposoby modelowania bayesowskiego, takie jak teoria detekcji sygnału (Green, Sweets, 1988), modelowanie probabilistyczne (Oaksford, 2001) czy koncepcja Thomasa L. Griffithsa, Charlesa Kempa i Joshui B. Tenenbauma (2008). Por. także Zednik, Jäkel, 2016.

możliwe dzięki zastosowaniu reguły Bayesa[15] (por. Anderson, 1990, 1991; Oaksford, Chater, 2007)[16]. W tym podejściu można wyróżnić następujące kroki (Oaksford, Chater, 2007, s. 71–72):

1. Określa się cele systemu poznawczego.
2. Opracowuje się formalny model środowiska, do którego dany system jest przystosowany.
3. Przyjmuje się minimalne założenia dotyczące obliczeń.
4. Uzyskuje się optymalną funkcję zachowania, określaną w krokach 1–3 (wymaga to formalnej analizy z wykorzystaniem racjonalnych norm, takich jak teoria prawdopodobieństwa, logika lub teoria decyzji).
5. Bada się dowody empiryczne, aby sprawdzić, czy przewidywania dotyczące funkcji zachowania są potwierdzone.
6. Iteracyjnie powtarza się kroki 1–5 w celu udoskonalenia teorii.

Drugą ramą jest hierarchiczne kodowanie predykcyjne, określane w niniejszej pracy jako PP. Podstawowa różnica pomiędzy *analizą racjonalną* a tym ujęciem polega na tym, że ta pierwsza może być używana do formalnej charakterystyki percepcji i działania jako rodzajów optymalnego wnioskowania probabilistycznego, zaś PP służy do opracowania algorytmów, które faktycznie wykonują tego rodzaju wnioskowania[17]. Oznacza to, że algorytmy opracowane w ramach badań nad PP mogą być postrzegane jako potencjalne odpowiedzi na pytania o to, jak dany system poznawczy robi to, co robi, tzn. można je potraktować jako opis procesów funkcjonalnych, które przyczyniają się do zachowania tego systemu. Trudno w tym momencie określić, na co zwracają uwagę Fink i Zednik, które z algorytmów faktycznie stanowią prawdziwy opis procesów funkcjonalnych w naszych mózgach, ale wielu autorów uważa odpowiedzi formułowane w ramach tego nurtu za główny wkład eksplanacyjny PP do badań nad ludzkim mózgiem (Fink, Zednik, 2017, s. 5). Zdaniem innych autorów (np. Gładziejewski, 2019; Harkness, Keshava, 2017; Hohwy, 2015a) PP może stanowić szkic mechanizmu, który pozwoli badaczom sformułować mechanistyczne wyjaśnienie określonych zjawisk poznawczych[18].

Można zatem stwierdzić, że rama badawcza PP jest bardziej interesująca i płodna z punktu widzenia zwolenników tezy o bayesowskiej naturze ośrodkowego układu nerwowego niż rama wyznaczana przez *analizę racjonalną*, ponieważ nie tyle postuluje ona, że przetwarzanie informacji w mózgu ma charakter bayesowski, jak

[15] Niektórzy autorzy (por. Colombo, Seriès, 2012) twierdzą, że nie jest konieczne rozważanie, w jaki sposób mózg implementuje regułę Bayesa, aby wyjaśnić zachowania i poznanie na modłę *analizy racjonalnej*. Można tworzyć biologicznie możliwe algorytmy, które aproksymują optymalne wnioskowanie probabilistyczne. Można w tym celu wykorzystać przykładowo nauczanie maszynowe, abstrahując tym samym od reguły Bayesa (por. Griffiths, Lieder, Goodman, 2015; Kwisthout, van Rooij, 2013).

[16] Krytykę tego podejścia przeprowadził m.in. Herbert Simon (1983) oraz Bowers i Davis (2012).

[17] Trzeba jednak dodać, że w sposób przybliżony, a nie dokładny.

[18] Tematem wyjaśniania mechanistycznego zajmuję się w rozdziale *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego*.

czyni się w tamtym podejściu, ile wskazuje pewne możliwe algorytmy dotyczące tego, w jaki sposób sieci bayesowskie są zaimplementowane w mózgu. Argument ten zyskuje na sile, kiedy powiąże się go z wyróżnionymi przez Marra poziomami opisu systemu przetwarzania informacji (1982).

2.4. Poziomy analizy a przetwarzanie predykcyjne

Marr stwierdził, że każdy system poznawczy może być analizowany na trzech różnych poziomach. Wyróżnił poziom obliczeniowy, algorytmiczny oraz implementacji. Na poziomie obliczeniowym określa się to, *co* robi dany system i *dlaczego* to robi. Innymi słowy, wskazuje się na funkcje realizowane przez dany system czy mechanizm. Na pytanie o *co* można odpowiedzieć, opisując zachowanie danego systemu. W odniesieniu do interesującego nas PP chodziłoby o określenie funkcji przetwarzania informacji, która mapuje informacje pochodzące z wejść zmysłowych na określone zachowania (por. Zednik, 2018, s. 389). Na tym poziomie określa się cele obliczeń, wskazuje odpowiednie reguły konieczne dla realizacji danego zadania oraz określa strategię przetwarzania danych. Pytanie o *dlaczego* dotyczy, czegoś, co Marr określa jako „trafność" (*appropriateness*) danego zachowania w odniesieniu do wykonywanego zadania (Marr, 1982, s. 24). W ujęciu Orona Shagrira (2010) oznacza to, że dane zachowanie może zostać uznane za „trafne" czy właściwe, jeżeli matematyczny opis tego zachowania może zostać zmapowany na (faktyczną lub kontrfaktyczną) właściwość środowiska, w którym działa dany system. Zatem nie da się wyjaśnić zachowania czy funkcji danego systemu, jeżeli nie uwzględni się środowiska, w którym dane zachowanie czy działanie jest realizowane. Uwagę tę powinniśmy zapamiętać[19].

Poziom algorytmiczny daje odpowiedź na pytanie dotyczące tego, w *jaki* sposób dany system (mechanizm) robi to, co robi. Określa się tutaj dane przetwarzania, dane na wejściu i wyjściu oraz wskazuje na sposób realizacji obliczeń. Modelowanie procesów poznawczych pozwala wyjaśnić dane zjawisko na tym poziomie. Model jest satysfakcjonujący eksplanacyjnie wtedy, gdy (1) jego zmienne odpowiadają komponentom, działaniom, właściwościom i organizacji wyjaśnianego mechanizmu, który jest odpowiedzialny za realizację danego zachowania czy funkcji; oraz (2) zależności pomiędzy zmiennymi w modelu odpowiadają relacjom przyczynowym zachodzącym pomiędzy komponentami wyjaśnianego mechanizmu (Kaplan, Craver, 2011, s. 611). Teorie algorytmiczne nie wyjaśniają jednak tego, jak i gdzie fizycznie realizowany jest dany system. To jest zadanie teorii budowanych na poziomie implementacji.

[19] Por. na ten temat podrozdziały *Wyjaśnienie mechanizmów normatywnych a ograniczenia* oraz *Normatywne mechanizmy predykcyjne a ograniczenia*.

Poziom implementacji wskazuje na miejsce, *gdzie* dany system jest fizycznie czy biologicznie realizowany, i sposób, w *jaki* to się dzieje. Wskazuje się tutaj np. na określony mechanizm neuronalny lub strukturę krzemową (w przypadku mechanizmów realizowanych przez procesor komputerowy), odpowiedzialny za wykonywanie takiej, a nie innej funkcji, oraz jego komponenty, własności i strukturalną organizację. Zdaniem Marra pełne wyjaśnienie danego zjawiska zakłada jego przeanalizowanie i objaśnienie na każdym z trzech omówionych poziomów (Marr, 1982, s. 23).

Badacze zwracają uwagę na niejasność dotyczącą relacji pomiędzy poszczególnymi poziomami. Ogólnie twierdzi się, że relacja pomiędzy poziomami jest relacją realizacji. Każdy poziom powinien być postrzegany jako realizacja poziomu wcześniejszego. Poziom implementacji realizuje algorytmy opisane na poziomie algorytmicznym, określając mechanizmy realizujące dane algorytmy, zaś poziom algorytmiczny jest realizacją poziomu obliczeniowego. Marr podkreślał jednak, że z innej strony poziomy są od siebie formalnie niezależne. Przykładowo wiele różnych algorytmów może realizować dany problem obliczeniowy, zaś algorytmy mogą być w różny sposób realizowane fizycznie.

Kolejny problem dotyczy sposobu, w jaki chce się wyjaśniać dane zjawisko. Zdarza się, że badacze nie potrafią wyjaśnić określonego zjawiska na każdym z trzech opisanych przez Marra poziomów. Zwolennicy podejścia „oddolnego" (*bottom-up*) (por. McClelland i in., 2010) sugerują, że sama znajomość algorytmów albo implementacji wystarczy, aby wyjaśnić dane zachowanie czy funkcję systemu. Innego zdania są zwolennicy podejścia odgórnego (*top-down*), do nich należy zaliczyć także Marra, którzy twierdzą, że opisanie zachowania na poziomie obliczeniowym jest równie istotne z punktu widzenia jego wyjaśnienia, jak uwzględnienie pozostałych dwóch poziomów. Łatwiej zrozumieć algorytm i implementację danego zachowania czy funkcji, jeżeli ujmie się mechanizm, który je realizuje i ucieleśnia (Marr, 1982, s. 27).

Powszechnie uważa się, że modele bayesowskie zlokalizowane są na poziomie obliczeniowym, ponieważ umożliwiają zrozumienie, dlaczego dany system robi to, co robi. Z tej perspektywy można zasadnie sądzić, że większą moc eksplanacyjną będą miały te modele, które umożliwią badaczom wgląd w inne poziomy analizy niż obliczeniowy. Czy tak jest w przypadku *analizy racjonalnej*? Zdaniem niektórych autorów (por. np. Colombo, Seriès, 2012) samo opisanie zachowania i poznania w terminach optymalnego wnioskowania probabilistycznego nie wymaga, a właściwie nawet nie sugeruje, że mózg faktycznie używa reguły Bayesa czy realizuje ją do obliczania rozkładu prawdopodobieństwa. Jeżeli tak jest, to niejasne pozostaje pytanie o to, czy *analiza racjonalna* pozwala na pełne (tzw. ujmujące wszystkie trzy wyróżnione przez Marra poziomy) wyjaśnienie badanych zjawisk.

W przypadku badań prowadzonych w ramie PP sytuacja prezentuje się inaczej. W rozwijanym tutaj ujęciu wielopoziomowy i hierarchicznie zorganizowany model

generuje kolejne predykcje minimalizujące potencjalne błędy predykcyjne obecne na niższych poziomach hierarchii. Na każdym z poziomów ma być realizowana reguła Bayesa w celu reprezentowania świata i skutecznego w nim działania (por. Clark, 2013b; Friston, 2005). Oprócz opisów na poziomie obliczeniowym rozwijane są różnego rodzaju algorytmy, które faktycznie realizują regułę Bayesa. Pozwalają tym samym na stawianie odpowiedzi na pytania o *jak* danego mechanizmu, czyli w pewnym stopniu funkcjonalnie wyjaśniają dany system. Jak wcześniej wspomniałem, niejasna pozostaje kwestia neuronalnej implementacji proponowanych algorytmów, niemniej wielu autorów czyni wysiłki, by wskazać jej możliwość (por. Friston, 2005; Gordon i in., 2018; Rao, Ballard, 1999; Parr, Friston, 2018).

Podsumowując ten wątek, należy stwierdzić, że badania z zakresu PP pozwalają na bardziej satysfakcjonujące wyjaśnienie zjawisk niż ma to miejsce w *analizie racjonalnej*. Przetwarzanie predykcyjne (PP) pozwala na wyjaśnienie poznania jako pewnej całości na drodze wyjaśnienia mechanizmów, które są realizowane przez odpowiednie algorytmy zaimplementowane w określonym środowisku fizycznym czy biologicznym. Istnieje jednak trudny obliczeniowo problem, który jest bezpośrednio powiązany z ramą PP i który daje się w niej rozwiązać. Mianowicie: błędy predykcyjne biorą się z nieznajomości przyczyn pobudzeń zmysłowych. Jeden skutek może być wywołany przez wiele różnych przyczyn, a także jedna przyczyna, np. jakieś zjawisko obecne w środowisku, może wywołać wiele różnych skutków, tj. pobudzeń zmysłowych. Tradycyjne podejścia obliczeniowe nie zawsze potrafiły sobie poradzić z tym problemem. Inaczej rzecz ma się w przypadku PP. Model generatywny, stosując hierarchiczne wnioskowanie bayesowskie, umożliwia wykorzystanie przyczynowej struktury bodźców zmysłowych do dostarczania formalnych ograniczeń (*constraint*), które dotyczą odwzorowania przyczyn i skutków. Zatem zamiast próbować cofać się od skutków, tj. pobudzeń zmysłowych, do wywołujących ich przyczyn, bayesowski model generatywny próbuje przewidzieć, jaka informacja powinna zostać dostarczona przez wejścia zmysłowe, o ile ma być ona wywołana przez określone zjawiska ze środowiska (Hohwy, Roepstorff, Friston, 2008, s. 688). Takie rozwiązanie problemu pobudzeń zmysłowych pozwala mówić o oryginalności i nowatorstwie prezentowanego tutaj ujęcia.

2.5. Problemy z modelami bayesowskimi[20]

Bayesowska teoria konfirmacji spotkała się z wieloma zarzutami. Najważniejsze wiążą się z tzw. paradoksem kruków[21], problemem logicznej wszechwiedzy[22] albo nowym problemem indukcji Goodmana. Ostatniemu z wymienionych należy przyjrzeć się bliżej. Paradoks ten ogólnie dotyczy faktu, że o możliwości zastosowania teorii konfirmacji przesądzają określone założenia filozoficzne, które należy wraz z nią przyjąć (w sposób jawny bądź nie). Bowiem jeżeli nie oprze się teorii konfirmacji na określonych założeniach filozoficznych (np. dotyczących sposobu konstruowania przestrzeni probabilistycznej albo jej przedmiotów), to prowadzi ona do absurdalnych wniosków, np. potwierdzających zdania ze sobą sprzeczne. Metodologicznie problem Goodmana można przedstawić następująco (Kawalec, 2003, s. 124):

> Dla dowolnej kwantytatywnej teorii konfirmacji T istnieją dwie (wzajemnie przekładalne) interpretacje I_1 oraz I_2, tak że dla zdania $\{\tau(\alpha) = x\} \in T$ w I_1 oraz I_2 stopnie konfirmacji tego zdania są paradoksalnie rozbieżne (w szczególności $\{\tau(I_1(\alpha)) = x\} \wedge \{\tau(I_2(\neg\alpha)) = x\}$, gdzie τ jest funkcją konfirmacji w T oraz $x \in [0, 1]$).

Paradoks Goodmana dowodzi, że nie da się skonstruować dobrej teorii konfirmacji bez przyjęcia minimalnych założeń filozoficznych. Założenia te dotyczyć mają m.in. rodzajów przedmiotów dopuszczalnych w danej teorii. Założenia jednak, na co zwraca uwagę Paweł Kawalec, rozumiane mogą być na tyle szeroko, aby nie prowadzić do określonych implikacji dla sporu realizm – antyrealizm (Kawalec, 2003, s. 124)[23]. Założenia te dotyczą m.in. następujących zagadnień (Kawalec, 2003, s. 125):

---

[20] W niniejszym paragrafie skoncentruję się jedynie na wybranych problemach bayesizmu ze względu na ich teoretyczną ważność dla omawianych tutaj zagadnień. Z powodu ograniczeń nałożonych na niniejsze analizy wiele trudności musi zostać ominiętych. Szersze ich omówienie odnajdzie czytelnik m.in. w następujących pracach: Easwaran (2011); Kawalec (2003, 2012) czy Mortimer (1982).

[21] Paradoks ten wiąże się z obserwacją Carla G. Hempla, że jego teoria konfirmacji (która stanowi podstawę dla bayesowskiej teorii konfirmacji) pozwala przewidzieć, iż obserwacja czarnego kruka (zgodnie z oczekiwaniami) potwierdza hipotezę, że wszystkie kruki są czarne. Jednak obserwacja nieczarnego niekruka, również potwierdza tę hipotezę. Można bronić tej teorii, twierdząc, że w świecie realnym potwierdzanie oparte na obserwacji czarnych kruków jest o wiele pewniejsze niż to oparte na obserwacji nieczarnych niekruków. Por. Fitelson 2006.

[22] Zgodnie ze standardowymi aksjomatami teorii prawdopodobieństwa wydaje się, że każda prawda logiczna ma mieć prawdopodobieństwo równe jeden, czyli być pewna. Użycie zwrotu „wydaje się" jest zasadne, ponieważ aksjomaty Kołmogorowa tak naprawdę nie pociągają za sobą tego twierdzenia. Jednak jeśli weźmie się pod uwagę normalną interpretację i stwierdzi, że zbiór możliwości jest jakimś zbiorem (logicznie lub metafizycznie) możliwych światów, oraz utożsami się zdania z zestawem światów, w których są prawdziwe, to oznacza to przypisanie podmiotowi to twierdzącemu logiczną (a być może nawet metafizyczną) wszechwiedzę. (Problem ten wiąże się z omawianym w różnych miejscach tej pracy problemem praktycznej obliczalności modeli bayesowskich). Zwolennicy bayesizmu odpowiadają na ten zarzut, twierdząc, że nie jest to teoria deskryptywna, a normatywna, w związku z czym są możliwe pewne idealizacje.

Inne ważne ograniczenia bayesizmu wiążą się m.in. z problemem prawdopodobieństwa pierwotnego czy nieadekwatności historycznej (por. Easwaran, 2011; Kawalec, 2012).

[23] Pewną interpretację owych przedmiotów przedstawiłem w rozdziale *Pojęcie przetwarzania predykcyjnego*, pisząc o rzeczywistych wzorcach. Por. podrozdziały *Rzeczywiste wzorce w przetwarzaniu predykcyjnym – S-reprezentacje i realne wzorce*.

- sposobu konstruowania przestrzeni probabilistycznej;
- przypisywania prawdopodobieństw pierwotnych hipotezom;
- przypisywania prawdopodobieństw warunkowych hipotezom;
- zasady uaktualniania prawdopodobieństwa hipotez itd.

W świetle tych uwag widoczne jest, że stanowisko bayesowskiego PP przyjmuje wiele założeń filozoficznych (m.in. dotyczących nieświadomego charakteru realizacji reguły Bayesa, natury predykcji czy informacji pochodzących ze środowiska). W PP aktualizacja przekonań wiąże się bezpośrednio z przypisywaniem określonej wagi błędowi predykcyjnemu. Waga opiera się na dokładnościach rozkładów prawdopodobieństwa (lub w przypadku ciągłym funkcji gęstości prawdopodobieństwa), gdzie precyzja jest odwrotnością wariancji, tak że wysoce precyzyjny rozkład ma bardzo małą wariancję. Waga wyraża prędkość uczenia się, która zwiększa się wraz z przypisywaniem precyzji danemu przekonaniu. Można tutaj mówić o mechanizmie optymalizacji precyzji. To znaczy: model może generować malejące przewidywania precyzji w celu stworzenia zmiennej szybkości uczenia się, która jest wrażliwa na informacje kontekstowe. Model generatywny w związku z tym musi być hierarchiczny, tak aby ciężar błędów predykcyjnych na dowolnym poziomie mógł być modulowany przez wyuczone prawidłowości w kilku zakresach czasoprzestrzennych, tj. na innych poziomach. Przykładowo: waga błędu predykcyjnego związanego z postrzeganiem samochodów poruszających się po ulicy zmniejszy się, gdy będę prowadził samochód w okularach (zakładając, że mam słaby wzrok) i gdy np. będę prowadził samochód po drodze, którą znam (Hohwy, 2020a, s. 2–3).

Należy jednak stwierdzić, że zwolennicy PP są dalecy od satysfakcjonującej eksplikacji wielu podstawowych założeń filozoficznych swoich ujęć[24] (por. Colombo, Elkin, Hartmann, 2018), co jest szczególnie wyraźne, kiedy weźmie się pod uwagę różnice pomiędzy konserwatywnym a radykalnym PP (oraz ich rodzajami). W niniejszej pracy odnoszę się tylko do niektórych z tych założeń, podkreślając jednocześnie, że wiele przyjętych przeze mnie rozwiązań albo wymaga dalszych, pogłębionych analiz, albo stanowi pewną idealizację.

Inne problemy wiążą się z tzw. dogmatyzmem bayesizmu, który polega na tym, że w punkcie wyjścia pewne możliwości się wyklucza, ponieważ niemożliwe stałoby się wyznaczanie prawdopodobieństwa branych pod uwagę możliwości przy wykorzystaniu twierdzenia Bayesa (Kawalec, 2012).

Zwolennicy PP mogą bronić się przed powyższymi zarzutami, twierdząc, że (1) model generatywny implementuje sieć bayesowską w sposób przybliżony, a nie w sposób dokładny (por. Hohwy, 2020a; Spratling, 2017); (2) wiele z argumentów krytykujących bayesizm odwołuje się do użycia teorii prawdopodobieństwa na

[24] Myślę, że jest to jeden z głównych powodów, dla którego PP należy uznać za pewną tradycję badawczą (w sensie Laudana) czy wręcz zbiorczą nazwę dla wielu różnych ujęć przyjmujących względnie podobne założenia aniżeli określoną teorię czy paradygmat badań.

poziomie personalnym, zaś większa część literatury i badań z zakresu PP skupia się na procesach subpersonalnych; (3) pojęcie prawdopodobieństwa w PP ma również charakter frekwencyjny, tzn. można tu mówić o pewnej „obiektywnej" strukturze probabilistycznej mierzonej w *statystykach scen naturalnych* (por. Orlandi, 2016). Takie rozwiązanie pozwala tłumaczyć, dlaczego pewne możliwości są w PP z góry wykluczane. Dzieje się tak ze względu na to, że świat jest w pewien sposób uprzednio ustrukturyzowany (por. podrozdział *Rzeczywiste wzorce w przetwarzaniu predykcyjnym*); można twierdzić, że w niektórych rodzajach operacji poznawczych mózg nie przeprowadza w pełni rozwiniętych i poprawnych formalnie złożonych obliczeń, ale wykorzystuje uproszczone heurystyki, które – jak potwierdzają badania empiryczne – mogą być w pewnych okolicznościach skuteczniejsze niż stosowanie złożonych obliczeń (por. Gigerenzer, Brighton, 2009; Parpart, Jones, Love, 2018). Taką możliwość wykazują niektórzy zwolennicy ramy aktywnego wnioskowania (por. Constant i in., 2019; Veissière i in., 2020). Trzeba jednak podkreślić, że takie próby są nieobecne w literaturze dotyczącej PP (por. Colombo, Elkin, Hartmann, 2018).

2.5.1. Problemy eksplanacyjne bayesizmu

Najmocniejsze argumenty przeciwko zastosowaniu modeli bayesowskich w kognitywistyce i neuronaukach przedstawili Jeffrey S. Bowers oraz Colin J. Davis w głośnym artykule *Bayesian just-so stories in psychology and neuroscience* (2012). Ich zdaniem to nie nowoczesne modele bayesowskie pozwalają wyjaśnić wiele zagadnień związanych z wykonywaniem różnych zadań przez nasze mózgi, ale tradycyjne niebayesowskie ujęcia. Dzieje się tak, ponieważ (1) modele bayesowskie wykorzystywane w psychologii mają słabe oparcie w dowodach empirycznych. Przekłada się to bezpośrednio na to, że modele te znacznie słabiej przewidują zjawiska niż modele klasyczne; (2) zastosowanie modeli bayesowskich w neuronaukach jest jeszcze słabiej uzasadnione empirycznie. Istnieje wiele matematycznych analiz, które wykazują, jak określone populacje neuronów mogłyby dokonywać obliczeń w sposób bayesowski, niemniej istnieje bardzo niewiele dowodów, które potwierdzają empirycznie te analizy; (3) nie jest dobrze uzasadnione przekonanie, że teorie psychologiczne powinny być ograniczane tylko lub przede wszystkim przez racjonalną analizę tego, co mózg powinien robić; (4) modele bayesowskie oraz ich założenia są zwykle dopasowywane *post hoc* do określonych danych (por. Litwin, Miłkowski, 2020). Zdaniem Bowersa i Davisa nie można się z bayesizmem zgodzić, bowiem istnieje wiele innych równie ważnych ograniczeń. Zaliczyć można do nich ograniczenia ewolucyjne, biologiczne czy obliczeniowe. Ignorowanie tych ograniczeń prowadzi do przekonania leżącego u podstaw modeli bayesowskich,

że działanie mózgu opiera się na optymalizacji zachowań skorelowanej z określonymi zadaniami. Innymi słowy, można mówić tutaj o zbytniej intelektualizacji dotyczącej pracy mózgu[25]. Powyższe uwagi po części, jak sądzę, można łączyć z tym, o czym mówiliśmy wcześniej. Mianowicie: modele bayesowskie są rozwijane przede wszystkim na obliczeniowym poziomie analizy wykazanym przez Marra. Inaczej jest, jak przekonują niektórzy, w przypadku modelowania bayesowskiego wykorzystywanego w PP. Czy zatem krytyka Bowersa i Davisa nie odnosi się do podejść opartych na PP? Trudno na to pytanie teraz jednoznacznie odpowiedzieć. Wrócimy do niego w dalszej części naszych analiz[26].

Inni badacze niejako rozwijają ostatni z argumentów autorów *Bayesian just-so stories in psychology and neuroscience*. Mianowicie: jeżeli spojrzy się na procesy poznawcze w kontekście ewolucji biologicznej gatunku ludzkiego, to posługiwanie się prawdopodobieństwem czy wprost regułą Bayesa jest dla człowieka czymś nietypowym, co pojawiło się stosunkowo późno w procesie ewolucji (por. Cosmides, Tooby, 1996; Tyszka, Zaleśkiewicz, 2001, s. 210).

Podane argumenty są mocne i ich przyjęcie może prowadzić do konkluzji, która odmawia mocy eksplanacyjnej części modeli bayesowskich, a być może nawet im wszystkim. Analizując je, należy zwrócić uwagę na fakt, że w większości podejść bayesowskich tworzy się modele idealnego obserwatora. Jak twierdzą David C. Knill i Alexandre Pouget (2004, s. 712) w modelach bayesowskich „ludzcy obserwatorzy zachowują się jak optymalni obserwatorzy bayesowscy". Wydaje się, że w podejściu predykcyjnym, dzięki odniesieniu do poziomów algorytmicznego i implementacji, ów idealny obserwator przestaje być idealny. Dzieje się tak, ponieważ modele w PP zorientowane są na faktyczne procesy, dzięki czemu odnoszą się one do określonych fizycznych komponentów oraz relacji przyczynowych, które są odpowiedzialne za realizację takich, a nie innych ludzkich zachowań (Harkness, Keshava, 2017). Zdaniem Matteo Colombo i Peggy Seriès (2012, s. 700) hipoteza bayesowskiego mózgu ma dwojaki charakter:

---

[25] Inny zarzut, o którym należy wspomnieć, wiąże się z tzw. problemem NP-trudnym (*NP-hard*). Dotyczy on praktycznej obliczalności modeli bayesowskich. Oznacza to, że wymogi bayesizmu, które powinien spełniać podmiot, są niemożliwe do zrealizowania. Dobrą ilustrację tego problemu przedstawia Kawalec (2006, s. 371–378). W kontekście PP por. Kwisthout, van Rooij, 2019. Możliwe strategie rozwiązania tego problemu omawiają van Rooij i in. 2018. Badacze ci jednak twierdzą, że żadna z tych strategii ostatecznie nie rozwiązuje problemu NP-trudnego. Pewną strategię rozwiązania tego problemu proponuje Chris Thornton (2016), twierdząc, że model generatywny aproksymujący wnioskowanie bayesowskie może implementować uniwersalną maszynę Turinga.

[26] Clark Glymour (2001) podjął próbę obrony implementacji algorytmów bayesowskich, twierdząc, że sieci bayesowskie to gatunek sieci neuronowych, które z kolei są gatunkiem graficznych modeli przyczynowych. Swoje twierdzenie opiera na obserwacjach i badaniach nad osobami z uszkodzeniami mózgu. Glymour zakłada, że wiedza na temat nieprawidłowego wykonywania określonych zadań przez pacjentów dostarcza informacji na temat struktury umysłu. Przykładowo uszkodzenie w acyklicznej sieci neuronowej powoduje pojawienie się nowych probabilistycznych relacji niezależności, co oznacza, że hipoteza bayesowskiej sieci neuronowej umożliwia testowalne przewidywania dotyczące uszkodzonego mózgu. Istotna dla analiz Glymoura jest możliwość graficznego modelowania procesów realizowanych w mózgu [pomysł ten rozwija uczeń Glymoura – David Danks (2014)].

1. Mózg wykonuje wnioskowania bayesowskie, dzięki czemu możemy sądzić i kierować swoimi działaniami w środowisku.

2. Mózg reprezentuje informacje zmysłowe pod postacią rozkładów prawdopodobieństwa.

Dla pełnej charakterystyki bayesowskiego wymiaru PP istotne jest, wprowadzone przez Jonesa i Love'a (2011), odróżnienie pomiędzy bayesowskim fundamentalizmem i bayesowskim instrumentalizmem, określanym przez nich jako oświecony bayesizm. Autorzy ci są zgodni co do tego, że modele bayesowskie pozwalają wyjaśnić sposób, w jaki poznanie wytwarza wzorce zachowań, niemniej ich zdaniem większość z tych modeli nie wyjaśnia tego, jak samo poznanie się odbywa. Dzieje się tak, ponieważ badacze ignorują wyjaśnienia z innych poziomów niż poziom obliczeniowy. Ściślej mówiąc: wyjaśnienia ludzkiego zachowania odwołujące się do indukcyjnego wnioskowania probabilistycznego (*inductive probabilistic inference*) są ich zdaniem niewystarczające. Takie podejście Jones i Love określają krytycznie mianem „bayesowskiego fundamentalizmu". Rama wyznaczana przez omawianą już wcześniej *analizę racjonalną* jest idealnym przykładem takiego podejścia. Charakterystyczne dla tego podejścia jest odwoływanie się do różnego rodzaju modeli generatywnych, często przybierających formę sieci przyczynowych, w których „prawdopodobieństwo zajścia odpowiednich zmiennych obserwacyjnych zależy od wartości nieobserwowalnych, ukrytych zmiennych. Hipotezy dotyczące obserwowalnych zmiennych odpowiadają wartościom zmiennych ukrytych" (Jones, Love, 2011, s. 175). Nie ma tutaj jednak mowy, co podkreślają Jones i Love, o faktycznych wyjaśnieniach psychologicznych, ponieważ podejście to nie mówi nic o statusie psychologicznym ukrytych zmiennych. Zmienne te służą jedynie do definiowania wspólnego rozkładu dla innych zmiennych, a model jest oceniany tylko w odniesieniu do tego, czy zachowanie człowieka jest zgodne z tym rozkładem. Dzieje się tak pomimo tego, że wielokrotnie mówi się tutaj o nieobserwowalnych konstrukcjach poznawczych, takich jak hipotezy i ich subiektywne prawdopodobieństwa[27]. Z tego powodu należy stwierdzić, że jest to bardzo ograniczona perspektywa, pomimo wielorakiego zastosowania. Głównej przyczyny jej ograniczenia można poszukiwać w zaniedbaniu analizy mechanizmów odpowiedzialnych za realizację określonych efektów, tj. działań, reprezentowania czy procesów podejmowania decyzji. Krytyka Bowersa i Davisa wydaje się zasadna w odniesieniu do tego podejścia. Inaczej jednak jest z bayesowskim instrumentalizmem.

W bayesowskim instrumentalizmie zwraca się uwagę na konieczność powiązania racjonalnego poziomu wyjaśniania (czyli poziomu obliczeniowego) z mechanistycznymi (tj. algorytmicznym i implementacji). Powiązanie to ma pierwszorzędne znaczenie, ponieważ, przykładowo, pozwala ono badaczowi wybierać odpowiednie

[27] Z tego powodów Jones i Love traktują bayesowski fundamentalizm jako nową formę behawioryzmu (por. Jones, Love, 2011, s. 175–176).

hipotezy badawcze poprzez odwołanie się do danych empirycznych. Innymi słowy, wybrane hipotezy można testować empirycznie (por. Jones, Love, 2011, s. 187). Nasze wcześniejsze analizy pozwalają przypuszczać, że PP jest modelem w stylu bayesowskiego instrumentalizmu. Ma ono mieć bowiem zastosowanie zarówno na poziomie analizy obliczeniowej, jak i mechanistycznej. Twierdzenie to jednak należy mocniej uzasadnić. Z tego powodu (1) bliżej przyjrzę się konstatacji, która głosi, że PP jest zastosowaniem inżynierii odwrotnej; oraz (2) zastanowię się nad tym, czy krytyka bayesowskiego PP jest zasadna w odniesieniu do zarzutu dotyczącego stosowania indukcyjnych wnioskowań probabilistycznych (por. James, Love, 2011, s. 170).

2.5.2. Inżynieria odwrotna w przetwarzaniu predykcyjnym

Elliot J. Chikofsky i James H. Cross przez inżynierię odwrotną rozumieją „proces analizy danego systemu w celu identyfikacji jego elementów i ich wzajemnych zależności oraz tworzenia jego reprezentacji w innej formie lub na wyższym poziomie abstrakcji" (Chikofsky, Cross, 1990, s. 15). Analizując powyższe określenie, nietrudno zauważyć, że inżynieria odwrotna jest metodą szeroko stosowaną w naukach zajmujących się złożonymi systemami. Stwierdzenie zatem, że kognitywistyka i PP posługują się tą metodą, wydaje się trywialne oraz mało interesujące poznawczo. Propozycję tę Miłkowski określa jako ujęcie szerokie i przeciwstawia mu ujęcie wąskie, które wiąże się z pojęciem przeprojektowania (*reenginering*), także zaproponowanym przez Chikofsky'ego i Crossa (Miłkowski, 2013, s. 13–14). Pojęcie to oznacza zarówno renowację (*renovation*), jak i rewizję (*reclamation*). Jest badaniem i zmianą danego systemu w celu odtworzenia go w nowej formie i późniejszej jego implementacji (Chikofsky, Cross, 1990, s. 15). W przeprojektowaniu chodzi zatem o rekonstrukcję albo replikację danego systemu w nową formę. To wąskie ujęcie jest znacznie bardziej dopasowane do zastosowania inżynierii odwrotnej w kognitywistyce. W naukach poznawczych, takich jak neurologia czy psychologia poznawcza, replikuje się dane zjawiska, mechanizmy czy systemy w różnych modelach. Można zatem powiedzieć, że kognitywistyka jest inżynierią odwrotną (Miłkowski, 2013, s. 14; por. podrozdział *Kant i przetwarzanie predykcyjne*). Przemawiają za tym następujące fakty: (1) nauki poznawcze wnioskują o danej funkcji na podstawie struktury; (2) wykorzystują one strategię dekompozycji w celu wyjaśnienia systemów poznawczych; (3) stosują inżynierię odwrotną i przeprojektowywanie w celu replikacji struktury systemów poznawczych i zrozumienia ich funkcji[28].

[28] „Wszystkie wyjaśnienia mechanistyczne rozpoczynają się od (a) identyfikacji zjawiska lub niektórych zjawisk, które mają być wyjaśnione, (b) przeprowadzenia dekompozycji na jednostki i działania istotne dla zjawiska, oraz (c) podania organizacji jednostek i działań, za pomocą których wytwarzają one zjawisko" (Illari, Williamson, 2012, s. 123).

Dla nauk poznawczych kluczowa jest zatem replikacja zjawisk. Jej pełne znaczenie ujawnia się wtedy, kiedy weźmie się pod uwagę wyjaśnienia mechanistyczne szeroko stosowane nie tylko w kognitywistyce, ale także naukach społecznych i inżynieryjnych (por. Bechtel, 2008; Craver, 2007; Craver, Darden, Machamer, 2011; Ylikoski, 2013, 2015). Replikacja danego zjawiska gwarantuje, że model tego zjawiska można uznać za kompletny. „Jedynym sposobem upewnienia się, że rozumiemy dany mechanizm i mamy jego pełny model przyczynowy, jest jego odtworzenie (replikacja) w innym medium" (Miłkowski, 2013, s. 19)[29].

Możemy, dokonawszy dodatkowych analiz, powrócić do naszego wyjściowego pytania dotyczącego tego, czy w PP stosuje się inżynierię odwrotną? Za pozytywną odpowiedzią na to pytanie przemawia następujący argument: PP jest jednym z rodzajów modelowania bayesowskiego. Powszechnie uważa się, że w modelach bayesowskich stosuje się inżynierię odwrotną (por. Colombo, Seriès, 2012; Tenenbaum i in., 2011; Zednik, Jäkel, 2014, 2016). Wiąże się to z przekonaniem, o czym już wspominałem, że modele bayesowskie mają zastosowanie na obliczeniowym poziomie analiz wyróżnionym przez Marra. Można zatem powiedzieć, że są one inżynieryjne w szerokim sensie inżynierii odwrotnej wyróżnionym przez Miłkowskiego. Istotne jest jednak to, czy są to modele replikacyjne, tj. czy można o nich mówić jako o inżynieryjnych w sensie wąskim? Sądzę, że tak. Modele wytwarzane w ramach PP nie tylko mają zastosowanie na poziomie obliczeniowym, ale – jak twierdzi wielu badaczy – także na poziomach algorytmicznym i implementacji. Sceptycy jednak podkreślają słabe zaplecze empiryczne PP oraz nieduże możliwości przewidywania zjawisk. Można oczywiście bronić PP w ten sposób, że odróżni się możliwości tego podejścia od ich faktycznego zastosowania, twierdząc, że w przypadku tej ramy badawczej za wcześnie jest na empiryczną weryfikację. Byłoby to jednak niesatysfakcjonujące poznawczo. Samo wykazanie, że PP potencjalnie ma zastosowanie na poziomach mechanistycznych, nie jest zadowalające. Argument ten należy wzmocnić.

Podsumowując ten wątek, trzeba powiedzieć, że PP potencjalnie można zaliczyć do bayesowskiego instrumentalizmu, ponieważ wykorzystuje ono inżynierię odwrotną jako metodę replikacji zjawisk[30]. Tym samym odróżniam to podejście od podejść fundamentalistycznych opierających się na szeroko rozumianej inżynierii odwrotnej.

Otwartą kwestią pozostaje pytanie o to, czy PP jako bayesowska inżynieria odwrotna prowadzi do wyjaśnień mechanistycznych? Trudno na to pytanie w tej chwili jednoznacznie odpowiedzieć. Jednym z powodów jest dość krótka historia PP i nieznajomość jego wszystkich możliwych zastosowań. Mamy raczej do czy-

[29] Miłkowski w tym samym miejscu dodaje: „Nie twierdzę, że zrozumienie mechanizmu jest zagwarantowane jego przeprojektowaniem. Ale pomaga ono stwierdzić, czy dany model jest kompletny, czy nie".

[30] Twierdzenie to zostanie lepiej uzasadnione w dalszych partiach niniejszej pracy.

nienia głównie z konstytucją teoretyczną i metodologiczną tego podejścia niż z jego empiryczną aplikacją. Jest to nowe i dynamicznie rozwijające się ujęcie i z tego powodu, o czym pisałem wcześniej, powinno się je raczej traktować jako ramę badawczą niż precyzyjny sposób modelowania różnego rodzaju czynności, np. psychicznych. Niemniej, niektórzy badacze uważają, że PP ma charakter mechanistyczny, jednak ich argumenty nie są oparte na mocnych empirycznych podstawach. Jedno jest jednak pewne: PP, w przeciwieństwie do innych modeli bayesowskich, pretenduje do bycia koncepcją wyjaśniającą, na każdym z wyróżnionych przez Marra poziomów, takich zjawisk jak percepcja, działania czy urojenia. Ta cecha czyni PP modelowaniem bayesowskim, które przynajmniej w jakimś stopniu jest odporne na zarzuty takich autorów jak Bowers i Davis czy Jones i Love. Do kwestii mechanistycznego statusu PP powrócę później.

2.6. Przetwarzanie predykcyjne jako bayesizm oświecony

Pewne niejasności budzi pytanie o to, czy PP pozwala na modelowanie zjawisk poznawczych na innych niż obliczeniowy poziomach opisu wyróżnionych przez Marra, tj. poziomach algorytmów i implementacji. Jeżeli chodzi o problem mechanistycznej implementacji, kwestia ta budzi wiele niejasności, o czym już kilkakrotnie wspominałem. Inaczej ona wygląda w przypadku zastosowania PP na poziomie algorytmicznym. Tutaj sytuacja jest dużo bardziej obiecująca. W niniejszym podrozdziale omówię kilka możliwych argumentów na rzecz tezy, że PP jest oświeconą metodą (w sensie Jonesa i Love'a) bayesowskiego modelowania czynności poznawczych. Pierwszym będzie argument z algorytmów. Następnie zwrócę uwagę na to, w jaki sposób w ramach PP należy rozumieć twierdzenie, według którego centralny system nerwowy konstruuje i testuje wewnętrzne modele świata zewnętrznego na drodze realizacji procesów poznawczych, które są aproksymacją wnioskowania bayesowskiego. Tym samym wykażę, wbrew krytykom, że w PP nie mamy do czynienia z wnioskowaniem indukcyjnym. Trzeci argument będzie dotyczył normatywności bayesowskiego modelu PP.

2.6.1. Argumenty za eksplanacyjnym charakterem bayesizmu

Argument z poziomu algorytmicznego

Przypomnijmy, że Spratling (2017; por. podrozdział *Możliwe algorytmy przetwarzania predykcyjnego i filozoficzne inspiracje*) wymienia pięć możliwych teorii,

które wykorzystują PP na poziomie algorytmicznym[31]. Są to liniowe kodowanie predykcyjne, kodowanie predykcyjne w siatkówce oka, kodowanie predykcyjne w pierwotnej korze mózgowej oparte na modelu Rao i Ballarda, kodowanie predykcyjne w pierwotnej korze mózgowej oparte na modelu model BC (*biased competition*) oraz kodowanie predykcyjne oparte na modelu FEP Fristona[32]. To, co jest wspólne wszystkim tym algorytmom, to obliczeniowy cel, jakim jest dopasowanie modelu do obieranych przezeń danych. Dopasowanie jest realizowane na drodze minimalizacji rozbieżności pomiędzy otrzymywanymi informacjami a parametrami modelu. Zbędne informacje są usuwane z sygnału przeznaczonego do dalszego przetwarzania, podczas gdy przewidywania (*estimates*) przyczyn pobudzeń zmysłowych są dalej przetwarzane. Omawiane algorytmy różnią się również tym, w jaki sposób wiążą współczynniki predykcyjne (*predictions coefficients*) z odpowiednimi biologicznymi obwodami neuronowymi. Otóż niektórzy uważają, że owe współczynniki odpowiadają ciężarom synaptycznym połączeń bocznych w siatkówce (por. Srinivasan, Laughlin, Dubs, 1982); inni wiążą je ze wskaźnikami wypalania komórek korowo-piramidalnych rzutujących sprzężenie zwrotne (por. Friston, 2009; Kiebel, Friston, 2011) lub z odpowiedziami odpowiednich sprzężonych komórek piramidalnych (por. Spratling, 2012). Kolejnym czynnikiem odróżniającym poszczególne algorytmy jest rola przypisywana procesom uczenia się przez model. Przykładowo w liniowym kodowaniu predykcyjnym uczenie się nie jest brane pod uwagę, podczas gdy w algorytmie opartym na FEP uczenie się, poprzez minimalizację wariancyjnej energii swobodnej, stosowane jest jako metoda określania parametrów modelu generatywnego (Spratling, 2017, s. 5).

Argument z wnioskowań percepcyjnych

Jones i Love twierdzą, że pomimo ogólnych założeń w poszczególnych modelach Bayesa, właściwie nie ma się do czynienia z zastosowywaniem jego reguły, ile z „matematycznie elegancką formalizacją abstrakcyjnej indukcji" (*mathematically elegant formalization of abstract induction*) (Jones, Love, 2011, s. 178). W praktyce oznacza to, że (1) model jest już wyposażony uprzednio w wiedzę, co przykładowo umożliwia mu rozpoznawanie wzorców (np. regularności w ciągu słów, w którym wymiary liter mają znaczenie), zatem fakt, że model rozpoznaje taki, a nie inny wzorzec nie jest zupełnie zaskakujący ani teoretycznie pouczający. Model, który ma wyjaśniać odpowiednią funkcję poznawczą danego systemu, zakłada sposób jej realizacji; oraz że (2) przyjmuje się tu mocne założenia dotyczące środowiska,

[31] Lista ta z pewnością nie jest kompletna.

[32] Modelem tym szczegółowo zajmuję się w rozdziałach *Zasada energii swobodnej w przetwarzaniu predykcyjnym* i *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego.*

w którym funkcjonuje dany system, tzn. tworzy się taki formalny model środowiska, do jakiego dany system ma być przystosowany. Obie te własności modeli w tym podejściu wiążą się z mocnym założeniem dotyczącym racjonalności wnioskowań obecnych w tradycyjnych modelach bayesowskich (por. Oaksford, Chater, 2007). Już klasyczna literatura dotycząca wnioskowań heurystycznych (por. Kahneman, Tversky, 1979; Kahneman, 2012) nie pozwala bezkrytycznie przejść wobec tego założenia. Mówienie o realizacji twierdzenia Bayesa jest zatem jedynie ogólną hipotezą przyjętą w tym podejściu.

Uwagi te nie są w pełni zasadne w odniesieniu do bayesowskiego PP. Mianowicie: w PP nie pojawiają się wnioskowania indukcyjne w sensie zdefiniowanym przez *analizę racjonalną*. Tam zachowanie danego organizmu jest rozumiane jako optymalne rozwiązanie określonego wnioskowania probabilistycznego, które dotyczy zadania wynikającego z kontaktu ze środowiskiem, na podstawie formalnego modelu środowiska. W PP, co dowodziłem w rozdziale *Pojęcie przetwarzania predykcyjnego*, wnioskowania nie są indukcyjne, ale abdukcyjne, przez co rozumiem, że są to wnioskowania do najlepszego wyjaśnienia (por. Seth, 2015). Hipotezy formułowane na podstawie tych wnioskowań najlepiej tłumaczą informacje, jeżeli traktuje się je jako dowody na rzecz twierdzeń, które dane hipotezy głoszą (por. Hempel, 1965, s. 372–374; Hohwy, 2014, s. 5). Oznacza to, że w PP system poznawczy „wyciąga" swoje przekonania (*priors*) z surowych danych w procesie narastającej optymalizacji modelu poprzez minimalizację błędów predykcyjnych. Model jest zatem konstruowany na drodze ciągłego uczenia się i minimalizacji błędów w ten sposób, że żadne określone przekonania nie muszą być z góry zakładane. Proces ten można nazwać „indukcyjnym" lub „abdukcyjnym", ale kluczowe jest to, że jest nienadzorowany: dane wejściowe nie są *a priori* sklasyfikowane, a przekonania w punkcie wyjścia mogą być ustalone losowo i potem stopniowo dopasowują się do statystyki danych wejściowych. Z tego powodu model generatywny można rozumieć jako samouwierzytelniający (*self-evidencing*) (Hohwy, 2014). Model w sensie ogólnym jest statystyczny i wielopoziomowy. Wychwytuje zależności, jakie zachodzą pomiędzy wartościami zbioru zmiennych losowych. Reprezentuje je za pomocą swoich wewnętrznych parametrów. Jego zmienne są ukryte, tzn. nie da się jedno-jednoznacznie przyporządkować odpowiednich danych do wytwarzających ich procesów przyczynowych. Innymi słowy, nie da się utworzyć takiego formalnego modelu świata, jaki się postuluje w innych modelach bayesowskich. Predykcje, które są generowane przez model generatywny, opisuje się w kategoriach rozkładu prawdopodobieństwa. Nie mamy tutaj do czynienia z formalnym modelem świata, ale z hierarchicznie zorganizowanymi rozkładami prawdopodobieństwa dotyczącymi odpowiednich informacji przekazywanych przez wejścia zmysłowe.

Powyższa charakterystyka jest czysto formalna, ponieważ nie wynika z niej, dlaczego i w jaki sposób model ma wykorzystywać wnioskowanie Bayesa oraz

dlaczego właśnie to wnioskowanie probabilistyczne, a nie jakieś inne (por. Colombo, Elin, Hartmann, 2018). Clark powiada (2013b, s. 189), że procesy poznawcze w ujęciu PP są aproksymacją wnioskowania bayesowskiego. Przez aproksymację należy rozumieć to, że dana funkcja albo proces są realizowane przy użyciu innej funkcji, która jest prostsza i łatwiejsza do zbadania i zastosowania. Takie wyjaśnienie niewiele mówi na temat tego, w jaki sposób mózg miałby „używać" reguły Bayesa[33]. Wydaje się, że z całą pewnością należy wykluczyć interpretację hiperrealistyczną, która dosłownie rozumie tutaj pojęcie aproksymacji, tzn. w tym sensie, że to mózg bezpośrednio wykorzystuje twierdzenie Bayesa. Należy również wykluczyć interpretację instrumentalną, czyli tę, która opiera się na uznaniu bayesowskiego charakteru wnioskowań w PP za pewien przydatny sposób mówienia (por. podrozdział *Wnioskowanie aktywne jako wnioskowanie enaktywne*).

Przekonująca jest propozycja Alexa Kiefera (2017), który proponuje, aby wyjaśnianie percepcyjne obecne w PP traktować dosłownie, czyli jako proces dochodzenia do prawdy i maksymalizacji koherencji modelu. Powołuje się on na pogląd Gilberta Harmana (1973), według którego wnioskowanie jest poszukiwaniem koherencji czy zmianą poglądów, co wiąże się z racjonalnym wymogiem utrzymywania prawdziwości w odniesieniu do odpowiednich reprezentacji lub przekonań. Argumentacja przyjęta przez Kiefera jest przekonująca, ponieważ autor ten pokazuje, w jaki sposób tak rozumiane wnioskowanie Bayesa można wykorzystać w modelach obliczeniowych, zaproponowanych np. przez Terrence'a J. Sejnowskiego i Geoffreya E. Hintona (1983), które były inspiracją dla współczesnych modeli obliczeniowych. Dopasowanie nowych informacji do modelu traktuje się jako szczególny przypadek uspójniania informacji lub – mówiąc inaczej – utrzymywania koherencji wewnętrznych parametrów modelu[34]. Charakterystyczne dla hierarchicznego modelu generatywnego w PP jest to, że wnioskowanie zależy tutaj od wielu skoordynowanych ze sobą poziomów. Zwiększenie koherencji modelu pozwala zatem na większą spójność w reprezentowaniu odpowiednich rozkładów prawdopodobieństwa. W PP precyzja (tj. wariancja odwrotna) (Clark, 2013b; Hohwy, 2013, rozdział 3) szacująca poziomy błędów predykcyjnych jest wykorzystywana do dookreślania już posiadanej wiedzy (*priors*) oraz nowych informacji docierających do modelu w odniesieniu do kolejnych poziomów hierarchii. Wnioski są precyzyjnie ważoną kombinacją uprzedniej wiedzy i prawdopodobieństwa pojawienia się określonej informacji. Precyzja zatem to sposób budowania kompromisu pomiędzy

[33] Gładziejewski proponuje, aby mówić tutaj o implementacji sieci bayesowskiej, której struktura przypomina sieć przyczynowo-probabilistyczną struktury środowiska (Gładziejewski, 2016, s. 571).

[34] Rozwiązaniu Kiefera można postawić następujący zarzut: poszukiwanie spójności jest problemem zasadniczo nierozstrzygalnym, a uspójnianie jest NP-trudne (chociaż nie NP-zupełne). Pojawia się tu zatem problem wszechwiedzy podmiotu oraz pytania o nieskończoną obliczalność modelu generatywnego. Zarzut ten można osłabić, twierdząc, że model generatywny wykorzystuje w obliczeniach uproszczone wnioskowania (heurystyki). Nie ma jednak jak dotąd takiej propozycji w literaturze poświęconej PP.

posiadaną wiedzą a przychodzącymi informacjami (Kiefer, 2017, s. 19). Tym sposobem model, wykorzystując nowe dane, zwiększa swoją wewnętrzną koherencję.

Argument z antyrealizmu

Ujęcie Kiefera stanowi nietrywialną interpretację tego, w jaki sposób rozumieć wnioskowania percepcyjne w PP. Nie jest ono jednak wolne od zarzutów. Mianowicie: nie pozwala ono odpowiedzieć na pytanie dotyczące tego, dlaczego to właśnie wnioskowanie Bayesa ma być tym wnioskowaniem probabilistycznym, które powinniśmy zastosować w algorytmach PP? Matteo Colombo, Lee Elkin i Stephan Hartmann (2018) wysuwają wiele obiekcji, które przemawiają na niekorzyść, podzielanego przez niektórych zwolenników PP, realistycznego podejścia w odniesieniu do wnioskowań bayesowskich. Po pierwsze podejście bayesowskie w PP i szerzej – w kognitywistyce nie realizuje specjalnych cnót epistemicznych, których nie podzielałyby inne alternatywne podejścia. Chodzi o to, że nie jest to podejście, które jest prostsze, mające większe możliwości unifikacyjne, czy przykładowo racjonalniejsze. Po drugie algorytmy bayesowskie mają jak dotąd słabe wsparcie w dowodach empirycznych. Zdaniem Colombo, Elkina i Hartmanna powinno się przyjąć podejście antyrealistyczne, co w praktyce oznacza agnostycyzm w odniesieniu do modeli bayesowskich[35].

Czy zatem wybór podejścia bayesowskiego w PP jest arbitralny? Istnieją powody, by tak sądzić. Można jednak podać kilka argumentów broniących zastosowania modeli bayesowskich, przynajmniej w nastawieniu antyrealistycznym. Część z nich formułują Colombo, Elkin i Hartmann. Ich zdaniem bayesizm wciąż stanowi najpopularniejsze podejście do badania niepewności i modelowania procesów poznawczych. Wybór tego podejścia jest więc posunięciem konserwatywnym, a konserwatyzm, jak wykazał P. Kyle Stanford (2019), sprzyjał rozwojowi specjalizacji w obrębie nauk empirycznych. Poszukiwanie nowych rozwiązań jest charakterystyczne dla współczesnej zinstytucjonalizowanej nauki, niemniej sam liberalizm badawczy nie jest wystarczającym argumentem za odrzuceniem konserwatywnego podejścia bayesowskiego. Statystycznie rzecz ujmując, bayesizm wciąż wiedzie prym w wielu badaniach nad niepewnością, sztuczną inteligencją czy procesami poznawczymi, zarówno w kognitywistyce, jak i innych dziedzinach wiedzy, takich jak ekonomia. Widać wyraźnie, że za wykorzystaniem podejścia bayesowskiego przemawia wiele argumentów instytucjonalno-socjologicznych. Z całą pewnością są to dobre argumenty na rzecz antyrealistycznego podejścia do bayesizmu. Niemniej,

[35] Wspomniani autorzy omawiają również inne, alternatywne metody modelowania niepewności (przykładowo tzw. prawdopodobieństwo kwantowe) (Colombo, Elkin, Hartmann, 2018, s. 12–18).

aby móc utrzymać podejście realistyczne (tak jakby chcieli przynajmniej niektórzy zwolennicy PP), potrzebny jest dodatkowy mocniejszy argument.

2.6.2. Normatywność bayesowskiego modelu przetwarzania predykcyjnego

Wielu badaczy podkreśla normatywny charakter wnioskowań opartych na regule Bayesa (por. Anderson, 1990; Hahn, 2014; Oaksford, Chater, 2007; Oaksford, 2014). Bayesizm jest normatywny w tym sensie, że ludzkie myślenie jest mierzone i oceniane w świetle reguł przezeń sformułowanych. Innymi słowy, w myśl tego poglądu modele wykorzystujące regułę Bayesa nie tylko opisują, w jaki sposób przebiegają procesy poznawcze i decyzyjne, ale także pokazują, w jaki sposób *powinny* one przebiegać. Dzięki temu możemy dane przekonania, decyzje czy działania określać jako trafne lub nie, poprawne bądź niepoprawne itd. Założenie o normatywności modeli bayesowskich ściśle wiąże się z założeniem dotyczącym racjonalności podmiotów, które myślą, podejmują decyzje i działają. Gdyby podmioty nie były racjonalne, to trudno byłoby nam wyjaśnić ich działania[36]. Stanowisko, które wiąże racjonalność z normatywnością, można określić jako normatywny racjonalizm (Elqayam, Evans, 2011, s. 235). Elqayam wraz z Evansem słusznie pokazują, że istnieją jednak inne koncepcje racjonalności, które ich zdaniem, nie mają charakteru normatywnego. Można za nimi wyróżnić następujące typy racjonalności:

- instrumentalna: zachowanie jest racjonalne, jeżeli polega na osiąganiu indywidualnych celów danego organizmu[37];
- ograniczona (*bounded*): zachowanie jest racjonalne, jeżeli respektuje ograniczenia poznawczo-biologiczne danego organizmu;
- ekologiczna: zachowanie jest racjonalne, jeżeli jest dostosowane do środowiska, w którym żyje i działa dany organizm;
- ewolucyjna: zachowanie jest racjonalne, jeżeli zostało ukształtowane przez ewolucję i służy celom wyznaczanym przez geny.

Zdaniem Elqayam i Evansa tylko racjonalność normatywna wiążę się z mocnymi założeniami preskryptywnymi. Można jednak, jak sądzę, zasadnie wątpić w to, czy pozostałe wyróżnione przez nich ujęcia nie są normatywne. Mianowicie: można przykładowo twierdzić, że w racjonalności instrumentalnej miarą bądź normą zachowania racjonalnego jest osiąganie indywidualnych celów. Jeżeli ta norma nie jest spełniona, to zachowania nie można określić jako „racjonalne". Dążenie do celu okazuje się miarą tego, co określa się tutaj jako „racjonalność". Analogicznie sprawa wygląda w kwestii pozostałych rodzajów racjonalności. Oczywiście zwolennik deskryptywizmu odpowie na to, że nie ma się tutaj do czynienia z powinnością. Nie

[36] Na konieczność przyjęcia tezy o racjonalności podmiotów zwraca uwagę Krystyna Bielecka (2018).
[37] Por. Bielecka, 2018 w sprawie obrony tego ujęcia.

jest bowiem tak, że dany osobnik *powinien* się zachowywać racjonalnie. Zgadzam się, jednak rozumienie normatywności przez pojęcie powinności jest jego znacznym zawężeniem. Wiele norm (np. w produkcji, w projektowaniu przedmiotów, ale także norm społecznych) czerpie swoją normatywność z tego, że dany przedmiot, wytwór czy zachowanie są z nimi zgodne bądź nie. Warto przytoczyć tutaj słowa Millikan (2011, s. 110), do których będę jeszcze wracał:

> pojęcia normatywne nie zawsze są oceniające. Pojęć normatywnych używa się do wskazania jakiegokolwiek rodzaju miary, od której są możliwe rzeczywiste odstępstwa. Na przykład średnia liczbowa stanowi normę jednego rodzaju, podobnie jak jakikolwiek rodzaj regularności: 'Przy takim niebie na zachodzie jutro powinno być słonecznie'.

I tak zasady *savoir-vivre'u* są pewnymi normami zachowywania się, niemniej nie dlatego, że zawierają imperatywy w stylu „Powinieneś tak używać sztućców" albo „Musisz w taki sposób ubierać się do pracy". Zasad tych można przestrzegać bądź nie. Są one pewną normą kulturową, której nieprzestrzeganie wpływa na ocenę naszych zachowań przez innych. Nie są to zatem normy w sensie prawnym albo moralnym z silnie zaznaczonym elementem powinnościowym[38] (na ten temat por. Piekarski, Wachowski, 2018, s. 401). Podobnie można rozumieć normatywność w innych rodzajach racjonalności niż tzw. normatywnej. Z tej perspektywy norma, jaką jest racjonalność, w niczym się nie różni od norm instrumentalnych, ekologicznych czy ewolucyjnych obecnych w tzw. nienormatywnych typach racjonalności. Ostatecznie te „nienormatywne" typy są tak samo normatywne jak racjonalność normatywna.

Powróćmy do Elqayam i Evansa. Ich zdaniem należy odróżnić te podejścia bayesowskie, które zakładają racjonalność normatywną, od tych, które zakładają racjonalności nienormatywne. Pierwsze podejście można określić mianem „preskryptywnego bayesizmu", podczas gdy pozostałe można nazwać „empirycznym bayesizmem" bądź „deskryptywizmem". Pierwsze stanowisko zakłada, że racjonalne myślenie, działanie czy procesy podejmowania decyzji *powinny* być mierzone względem prawdopodobieństwa bayesowskiego jako systemu normatywnego; drugie zaś mówi tylko tyle, że myślenie, działania czy decyzje *są* zgodne z prawdopodobieństwem Bayesa (Elqayam, Evans, 2011, s. 234). Podejście preskryptywne zakłada zatem możliwość przejścia pomiędzy „jest" (fakt empiryczny dotyczący natury myślenia, działania czy podejmowania decyzji) a „powinno być" (reguła, zgodnie z którą myślenie, działanie czy podejmowanie decyzji musi albo powinno reflektować regułę Bayesa; por. podrozdział *Normatywność homeostazy*). Twierdzę jednak, że jest to znaczne uproszczenie i zafałszowanie problemu, ponieważ pojęcie zgodności jest normatywne w znaczeniu, o którym wspomniałem wcześniej. Coś

[38] Por. podrozdział *Normatywność homeostazy*.

jest zgodne albo nie jest zgodne z twierdzeniem Bayesa. Oczywiście zgodność nie implikuje powinności, niemniej, aby ocenić coś jako myślenie, działanie czy podejmowanie decyzji itd., owa zgodność jest miarą ich bayesowskiego charakteru. Ta „zgodność" jest swego rodzaju modalnością normatywną. Nie jest to oczywiście modalność normatywna, jaką np. wykazuje modalność konieczności. Ta ostatnia, przyjmując zasadność analiz Elqayam i Evans, wiązałaby się z ujęciem preskryptywnym i przekonaniem, że myślenie, percepcja, procesy decyzyjne itd. muszą być zgodne z twierdzeniem Bayesa. Takie ujęcie określam mianem „mocnego preskryptywizmu", który przeciwstawiam słabemu preskryptywizmowi, a nie jak Elqayam i Evans – deskryptywizmowi. Deskryptywizm w tym ujęciu mówi tylko, że dane działanie, myślenie czy percepcja jest realizacją twierdzenia Bayesa. Nie mówi jednak nic o zgodności z tym twierdzeniem. W tej perspektywie za każdym razem, kiedy mówi się o bayesowskim charakterze pewnych procesów, funkcji czy mechanizmów psychicznych, zakłada się stanowisko deskryptywne. Jednak, gdy chce się wyjaśnić, dlaczego są one bayesowskie, należy przyjąć, któreś z podejść preskryptywnych. Oś sporu opisanego przez Elqayam i Evansa nie dotyczy zatem wyboru pomiędzy preskryptywizmem a deskryptywizmem, ale wyboru pomiędzy słabym a mocnym preskryptywizmem.

Część badaczy (por. Colombo, Elkin, Hartmann, 2018; Elqayam, Evans, 2011; Jones, Love, 2011) uważa, że w nauce, a przede wszystkim w psychologii i kognitywistyce, należy odrzucić podejście preskryptywne na rzecz empirycznego czy deskryptywnego. Dosadnie to wypowiadają Elqayam z Evansen (2011, s. 248):

> Nasz argument polega na tym, że psychologia ludzkiego myślenia byłaby lepsza, gdybyśmy przyjęli program deskryptywny. Normatywizm odegrał ważną rolę w historii badań nad ludzkim myśleniem. Podejście deskryptywne postrzega badania teoretyczne nad rozumowaniem i badaniami nad procesami decyzyjnymi jako opisowe, a nie oceniające. Celem nie jest ocenianie zachowań, ale zrozumienie ich oraz przewidywanie przy użyciu wszystkich odpowiednich narzędzi teoretycznych i metodologicznych. Modele formalne i obliczeniowe mają tutaj do odegrania ważną rolę. Nie muszą być one używane do oceny lub uzasadniania wnioskowań w stylu „jest – powinno być". Twierdzimy, że rozważania ewaluacyjne należy przywoływać tylko w badaniach edukacyjnych i innych zastosowaniach, w których celem jest poprawa ludzkiego myślenia i działania. Odejście od normatywizmu w kierunku deskryptywizmu odegrało kluczową rolę w rozwoju językoznawstwa jako dojrzałej nauki. Podobna zmiana kierunku może okazać się równie korzystna dla badania ludzkiego rozumowania i podejmowania decyzji.

W kontekście tej dyskusji należy zadać dwa istotne pytania: (1) czy krytyka normatywnego bayesizmu jest zasadna również w odniesieniu do bayesowskiego PP?; (2) czy dla PP istotne jest pytanie o normatywność modelowania bayesowskiego, czy może raczej jest tak, że PP zakłada podejście deskryptywne? Ustosunkowanie

się do powyższych pytań jest o tyle istotne, o ile pozwoli nam dać pozytywną bądź negatywną odpowiedź na problem postawiony w tej części pracy. Mianowicie: czy PP można uznać za oświecony bayesizm, czyli zasadną i nietrywialną ramę badawczą dla wyjaśnienia funkcji poznawczych, kontroli działań, podejmowania decyzji itd.? Innymi słowy, czy uzasadniony jest realizm w odniesieniu do predykcyjnego podejścia bayesowskiego? Zgodnie z naszymi wcześniejszymi analizami wyjaśnienie czegoś w kategoriach twierdzenia Bayesa zakłada jakąś wersję preskryptywizmu. Pytanie zatem nie dotyczy tego, czy PP zakłada preskryptywizm, tylko tego, *jaki* jego rodzaj zakłada. Jeżeli zakłada mocny preskryptywizm, to zasadne jest przyjęcie stanowiska realizmu w odniesieniu do bayesowskiego PP. Jeżeli zaś zakłada słaby preskryptywizm, to z całą pewnością zasadne jest podejście antyrealistyczne. Zanim przejdę do odpowiedzi na powyższe pytania, konieczne jest rozwinięcie dodatkowych rozważań.

2.6.2.1. Normatywność a poziomy opisu

W kontekście dyskusji nad normatywnością modeli bayesowskich Elqayam i Evans (2011) zwracają uwagę na fakt, że jeżeli chce się opisać wiedzę, przekonania czy jakieś funkcje poznawcze, to odpowiednim pojęciem służącym temu celowi jest wprowadzone przez Noama Chomsky'ego pojęcie „kompetencja" (por. Chomsky, 1965). Kompetencja oznacza wyidealizowaną, abstrakcyjną zdolność traktowaną jako psychiczna lub umysłowa własność bądź funkcja. Oprócz kompetencji, którą można rozumieć jako „znajomość czegoś" (np. danego języka), Chomsky wyróżnił także tzw. wykonanie (*performance*), czyli praktyczne użycie danej wiedzy (np. języka) w konkretnej sytuacji. Elqayam i Evans (2011, s. 368) z rozważań nad badaniami Chomsky'ego wyciągają dwa istotne wnioski. Po pierwsze pojęcia kompetencji oraz wykonania nie mają charakteru normatywnego. Mianowicie: przeciwieństwem kompetencji nie jest niekompetencja, ale wykonanie. Oznacza to, że nie musimy wprowadzać normatywnej dystynkcji kompetencja/niekompetencja, ponieważ aby opisać faktyczną zdolność językową czy jakąś funkcję poznawczą danego osobnika, wystarczą nam „nienormatywne" pojęcia zaproponowane przez Chomsky'ego. Po drugie kompetencja w sensie Chomsky'ego ma być analogiczna do wyróżnionego przez Marra algorytmicznego poziomu opisu. Czy wnioski wyciągnięte przez tych badaczy są zasadne? Najpierw skoncentruję się na drugim z wymienionych wniosków.

Marr (1982) z poziomem algorytmicznym wiązał kwestie związane z odpowiedzią na pytania dotyczące tego, w *jaki* sposób dany system bądź mechanizm robi to, co robi. Istotny jest w tym kontekście fakt, że teorie budowane na podstawie tego poziomu, nie wyjaśniają tego, gdzie i jak fizycznie realizowany jest dany system bądź

mechanizm. Wyjaśnienie tej kwestii leży w gestii teorii budowanych na poziomie implementacji. Elqayam i Evans zauważają zasadnicze podobieństwo pomiędzy tak rozumianym poziomem algorytmów a pojęciem kompetencji w rozumieniu Chomsky'ego. Zatem jeżeli kompetencja jest opisem systemu bądź mechanizmu na poziomie algorytmów, to wykonanie jest opisem z poziomu obliczeniowego, czyli tego, który odpowiada na pytanie o to, co dany system bądź mechanizm robi. Zdaniem Elqayam i Evansa opisy formułowane przez teorie algorytmiczne i obliczeniowe zadają zasadniczo inne pytania badawcze. Przykładowo teoria algorytmiczna odnosi się do pytania o to, w jaki sposób podejmowana jest dana decyzja. Teoria normatywna, w ich ujęciu, zadaje pytania w stylu: „Jaki powinien być dobry użytek z negacji w języku?". Teorie obliczeniowe i algorytmiczne są teoriami deskryptywnymi, zaś teorie normatywne są preskryptywne i jako takie, zdaniem przywoływanych badaczy, są zupełnie zbędne w opisie i wyjaśnieniu odpowiednich zjawisk. Tym samym dotarliśmy do pierwszego wniosku sformułowanego przez Elqayam i Evansa: pojęcia wyjaśniające dane funkcje poznawcze nie są pojęciami normatywnymi. Uważam, że ten wniosek jest przyjęty zbyt pochopnie. Teraz uzasadnię moje przekonanie.

Elqayam z Evansem wykorzystują wczesne rozumienie kompetencji przez Chomsky'ego. Powołują się przy tym na jego *Aspects of the theory of syntax* wydane w 1965 roku. Znamienne jest to, że Chomsky swoje rozumienie kompetencji zmieniał kilkukrotnie. W latach 80. pojęcie kompetencji traci swój lingwistyczny charakter, związany z posiadaniem pewnej wiedzy na temat gramatyki języka, na rzecz znaczenia psychologicznego. Mianowicie: kompetencja jest teraz rozumiana w kategoriach struktur mentalnych. Znajomość języka polega bowiem na realizacji określonych stanów mentalnych składających się z pewnego systemu reguł i zasad. Należy zatem mówić nie tyle o posiadaniu pewnej wiedzy, ile o spełnianiu pewnych stanów umysłu bądź mózgu (por. Chomsky, 1986). Oczywiście tak rozumianej kompetencji jest znacznie bliżej do mechanistycznego poziomu implementacji niż do poziomu algorytmów. Trudno jest zatem zgodzić się z konstatacją Elqayam i Evansa, że analizy z poziomu implementacji „nie są istotne dla prowadzonej tutaj dyskusji" (Elqayam, Evans, 2011, s. 239). Są istotne z dwóch powodów. Pierwszy wiąże się ze znaczeniem pojęcia kompetencji, o czym przed chwilą wspomniałem, drugi odnosi się zaś do problemu normatywności. Przejdę teraz do niego.

Pojęcia normatywne Elqayam i Evans rozumieją w dość prosty i jednowymiarowy sposób, wiążąc je z takimi dystynkcjami jak dobry/zły czy prawidłowy/nieprawidłowy. Przykłady, na które się powołują, nie tyle są błędne, ile trywialne. Jak mamy np. rozumieć podane już wcześniej pytanie o „dobry użytek z negacji w języku"? Inne przykłady proponowane przez tych autorów wiążą się ze zjawiskiem hazardu, który uważany jest za irracjonalny. Elqayam i Evans sugerują, że zwolennicy normatywizmu ową irracjonalność chcieliby wiązać z „niemoralnością" czy czymś w tym stylu. Jest to ewidentne nadużycie. Chodzi o to, że przeciwnik

jest przez nich bardziej projektowany niż rekonstruowany. Normatywność nie wiąże się bezpośrednio z moralnością. To raczej moralność jest pewnym podtypem normatywności[39]. Elqayam i Evans, odrzucając *quasi*-moralne oceny, jednocześnie odrzucają wszelkie normatywne sądy i kwantyfikacje. Innymi słowy, stwierdzają, że pojęcia, za pomocą których opisujemy i wyjaśniamy nasze praktyki czy funkcje poznawcze, są pozbawione jakichkolwiek treści normatywnych. Przykładem tego ma być „nienormatywna" para pojęć kompetencja/wykonanie. Twierdzę, że jest to fundamentalny błąd.

Aby uzasadnić moje przekonanie, omówię prosty przykład. Wyobraźmy sobie, że Jan wypowiedział następujący zwrot: „Ola osa". Jak wyjaśnić komuś, że jest to zdanie niepoprawne, nie używając pojęć normatywnych? Spróbujmy to zrobić. Możemy założyć, że Jan posiada odpowiednią kompetencję językową, zatem mamy problem z wykonaniem. Jak ten problem opisać, nie używając takich zwrotów jak: „niepoprawne", „błędne", „niewłaściwe"? Może być również tak, że wykonanie Jana jest bez zarzutu, tylko ma on „jakiś" problem z kompetencją. Można powiedzieć wtedy, że jego kompetencja jest niepełna, bo przykładowo nie wie on, że zdanie w języku polskim musi mieć strukturę podmiotu i orzeczenia. Może być też tak, że z jakiegoś powodu Jan jako dziecko przeżył groźny wypadek, w wyniku czego zaczął używać zdań pozbawionych orzeczenia. Jak opisać taką kompetencję bez użycia zwrotów normatywnych? Możemy spróbować mówić, że jest ona niepełna albo niezupełna. Pójdźmy jednak dalej. Jan wypowiada takie zdania, ponieważ stwierdzono u niego specyficzną formę afazji polegającą na nieużywaniu czasowników. Jak wyjaśnić afazję, którą w psychiatrii rozumie się jako pewną utratę zdolności do mówienia lub – jeszcze dosłowniej – jako dysfazję[40]. Czy zatem problem z wypowiadaniem zdań nie jest pewnego rodzaju dysfunkcją kompetencji? Zwolennik tezy Elqayam i Evansa mógłby jednak utrzymywać, że pojęcie dysfunkcji można zastąpić jakimś innym, nienormatywnym pojęciem i np. powoływać się, jak zresztą się czyni, na fakt, że afazja jest efektem uszkodzenia mózgu. Tym samym opuściłby on poziom algorytmiczny i przeszedł na poziom implementacji, który – przypomnijmy – należy wiązać z późniejszym rozumieniem kompetencji przez Chomsky'ego.

Zgodnie z główną tezą wywodów Elqayam i Evansa poziomy wyróżnione przez Marra nie mają żadnego związku z teoriami normatywnymi. Zadają bowiem inne pytania badawcze. Przekonanie to również budzi poważne wątpliwości. Od lat 80. tacy badacze jak Millikan, próbując wyjaśnić funkcje organizmów, stwierdzili, że bez pojęcia normatywności nie da się tego zrobić (por. także Bickhard, 2003, 2014; Millikan, 2011). Millikan przykładowo zarzucała (nienormatywnej) teorii funkcji

[39] Na ten temat zob. podrozdział *Normatywność homeostazy*.

[40] Wielu autorów zwraca uwagę na konieczność użycia słownika normatywnego w ocenie zaburzeń psychicznych. Por. Muders, 2014; Stier, 2013.

Roberta Cumminsa, że nie pozwala wytłumaczyć zaburzeń np. serca bądź żołądka[41]. Powszechnie uważa się, że wiele funkcji psychicznych czy poznawczych trzeba opisać, używając pojęć normatywnych. Przykładowo reprezentacje są normatywne w tym sensie, że można mówić o ich poprawności czy błędności, prawdziwości bądź fałszywości (por. Bickhard, 2014, 2015a, 2015b; Bielecka, 2018).

Mocny argument na rzecz powiązania mechanistycznego poziomu implementacji z teoriami normatywnymi podaje Justin Garson (2013). Podkreśla on, że każdy mechanizm realizuje odpowiednią funkcję bądź ich zbiór. Ich normatywność wywodzi się z normatywności funkcji. Zatem mechanizmy realizujące funkcje normatywne są w pewnym sensie mechanizmami normatywnymi. Jak to rozumieć? Pomyślmy sobie, że jeżeli *X* jest pewnym systemem, a *Y* jest aktywnością tego *X*, to *X* jest mechanizmem (lub jego częścią) dla *Y*, o ile *X* ma funkcję *Y*. Przykładowo: serce jest systemem, który pompuje krew. Serce jest zatem mechanizmem dla pompowania krwi, o ile ma ono tę funkcję. Jeżeli serce przestanie pompować krew, oznacza to, że jest ono dysfunkcyjne, czyli mechanizm odpowiedzialny za realizację tej funkcji przestał działać lub działa niepoprawnie. Dowodzi to tego, że nie istnieją jakieś dodatkowe mechanizmy „patologiczne", np. odpowiedzialne za atak serca, ponieważ atak serca nie jest jego funkcją. W takich przypadkach ma się do czynienia z „uszkodzeniem", „zepsuciem" czy „przerwaniem pracy" danego mechanizmu, a nie z pojawieniem się innego. Stwierdzenie zatem, że dany mechanizm jest normatywny, oznacza po prostu, że jest możliwe, aby dany system *X* był mechanizmem dla aktywności *Y*, pomimo faktu, że (np. w danym momencie) *X* nie może wykonać *Y*. Oznacza to, że o „uszkodzeniu" danego mechanizmu można mówić, kiedy nie jest on w stanie realizować danej funkcji. Brak możliwości realizacji danej funkcji jest warunkiem koniecznym dla „niedziałania" danego mechanizmu, nie jest jednak warunkiem wystarczającym (np. niepodłączony do prądu mikser nie może wykonywać swojej funkcji, co nie oznacza od razu, że jest on uszkodzony). Podsumowując, należy stwierdzić, że słownik normatywny („dysfunkcja", „patologia", „nienormalność" itd.) jest konieczny dla wyjaśnienia funkcjonowania mechanizmów i realizacji odpowiednich funkcji bądź ich braku. Wracając do Jana, należałoby powiedzieć, że jego problemy z wypowiadaniem poprawnych zdań w języku polskim wynikają z dysfunkcji odpowiedniego mechanizmu albo związanego z realizacją takiej, a nie innej kompetencji językowej, bądź z uszkodzeniem takiej, a nie innej części ośrodkowego układu nerwowego.

Wykorzystując uwagi Garsona, trzeba stwierdzić, że prawidłowe wyjaśnienie funkcjonowania oraz realizacji odpowiednich mechanizmów poznawczych i biologicznych wymaga odwołania się do pewnych podstawowych pojęć normatywnych, takich jak dysfunkcja czy patologia. Nie można się zatem zgodzić z sugestią

[41] Por. rozdział *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego*, szczególnie podrozdział *Funkcje w ujęciu Millikan*.

Elqayam i Evansa, że słownik normatywny jest zasadny tylko w przypadku, gdy używa się go w edukacji, rozwoju danej branży czy tworzeniu jakichś planów funkcjonowania, np. przedsiębiorstwa. Gdyby tak było, to z punktu widzenia lingwisty dziwne konstrukcje zdaniowe tworzone przez Jana nie byłyby zupełnie interesujące i musiałby je potraktować jako jakąś ekstrawagancję językową. Niepoprawny sposób mówienia Jana podlegałby jedynie ocenie jego nauczyciela, pracodawcy czy wspólnika. Konkludując: nie da się zupełnie odrzucić podejścia preskryptywnego w uprawianiu nauki, w tym kognitywistyki czy filozofii umysłu. Problematyczne jest jednak to, jak mocno należy interpretować preskryptywizm w odniesieniu do interesującego nas podejścia bayesowskiego. Należy zatem określić, w jakim stopniu wnioskowania bayesowskie są normatywne czy konstytutywne dla odpowiednich funkcji poznawczych. Innymi słowy, wiemy już, że stanowisko preskryptywizmu bayesowskiego jest *a priori* możliwe. Nie wiemy jednak, na ile jest ono zasadne i czy jest zasadne w odniesieniu do interesującego nas bayesowskiego PP. Temu zagadnieniu poświęcę dalsze analizy.

2.6.2.2. Preskryptywizm epistemiczny i ontyczny

Nasze wcześniejsze ustalenia pozwalają na wstępne uznanie tezy, zgodnie z którą zasadne jest używanie słownika normatywnego w odniesieniu do perspektywy bayesowskiej. Wiemy już, że istnieją mocne racje przemawiające za tym, że pojęcia normatywne mają uzasadnione znaczenia dla analiz przeprowadzanych na poziomach opisu wyróżnionych przez Marra. Niemniej, jak dotąd udało się wykazać tylko tyle, że istnieją pewne normatywne zależności w odniesieniu do opisu funkcjonowania mechanizmów czy realizacji określonych funkcji. Wciąż nie wiadomo, na ile sama struktura i funkcjonalność bayesowskiego modelu generatywnego jest normatywna dla percepcji czy działań podejmowanych przez dany organizm. Jeżeli uda się wykazać ową normatywną zależność, to zostanie wybroniony realizm w odniesieniu do perspektywy bayesowskiej w PP. Jeżeli zaś się nie uda, to wciąż możemy bronić antyrealistycznej interpretacji PP jako oświeconego podejścia bayesowskiego.

Wiemy już, że słaba wersja preskryptywizmu, która zakłada zgodność procesów, funkcji i mechanizmów psychicznych, jest zasadna w odniesieniu do PP. Pozwala ona przyjąć antyrealizm w odniesieniu do PP. Ów słaby preskryptywizm określę mianem „epistemicznego", ponieważ wiążę go z możliwością wyjaśniania zjawisk przy użyciu twierdzenia Bayesa. Pełne uzasadnienie realizmu w odniesieniu do bayesowskiego PP zakłada jednak przyjęcie mocnego preskryptywizmu, który od teraz będę nazywał „ontycznym". Ontyczny preskryptywizm zakłada, że zjawiska wyjaśnia się za pomocą twierdzenia Bayesa, ponieważ są one bayesowskie *per se*.

Jeżeli są one bayesowskie z natury, to znaczy, że *powinno* się je wyjaśniać przy użyciu modelowania Bayesa. Tym sposobem hipoteza dotycząca bayesowskiej natury charakteru PP uzasadnia hipotezę eksplanacyjną.

Analizy przeprowadzone w niniejszym rozdziale pozwalają zasadnie przyjąć hipotezę eksplanacyjną, która zakłada preskryptywizm epistemiczny i antyrealistyczne podejście do PP. W dalszych rozważaniach uzasadnię hipotezę dotyczącą natury (hipotezę ontyczną), tym samym wykazując zasadność preskryptywizmu ontycznego w odniesieniu do PP oraz przyjęcia stanowiska realistycznego. Niniejsze rozważania pozwalają również zasadnie bronić przekonania, zgodnie z którym bayesowskie modelowanie PP należy zaliczyć do tzw. oświeconego bayesizmu (nazywanego również „instrumentalizmem"), czyli stanowiska, które zakłada konieczność powiązania racjonalnego poziomu wyjaśniania (odnoszącego się do poziomu obliczeniowego) z poziomami mechanistycznymi (czyli algorytmów i implementacji). Powiązanie to jest konieczne, ponieważ pozwala ono badaczom wybierać odpowiednie hipotezy badawcze poprzez odwołanie się do danych empirycznych i tym samym umożliwia wyjaśnianie zjawisk poprzez odwoływanie się do faktycznych mechanizmów i ich organizacji (Jones, Love, 2011, s. 187).

Podsumowanie

W rozdziale tym argumentowałem za eksplanacyjnym charakterem bayesowskiego modelu PP. Pokazałem, że powszechna krytyka modeli bayesowskich jako opisujących poznanie i działanie tylko na obliczeniowym poziomie Marra, pomimo że trafnie charakteryzuje wiele modeli bayesowskich, np. *analizę racjonalną*, to nie jest zasadna w odniesieniu do PP. Stwierdziłem, że istnieje wiele różnych możliwych algorytmów, które pokazują, w jaki sposób PP modeluje czynności poznawcze, co w praktyce oznacza, że oferuje ono wyjaśnianie na poziomie algorytmicznym. Pokazałem także, że istnieją dobre argumenty na rzecz tezy, zgodnie z którą PP ma zastosowanie do mechanistycznego poziomu implementacji. W dalszej części argumentowałem, że bayesowskie PP należy uznać za stanowisko, które Jones i Love (2011) określili jako bayesizm oświecony, przeciwstawiając go słabemu eksplanacyjnie bayesowskiemu fundamentalizmowi.

W dalszej części tego rozdziału uzasadniałem tezę, zgodnie z którą PP zakłada pewną wersję preskryptywizmu. Przeprowadziłem dyskusję z głośnym twierdzeniem Elqayam i Evansa (2011), dotyczącą postulatu porzucenia myślenia o wyjaśnianiu poznania w kategoriach preskryptywnych i nakazującą przejść na pozycje deskryptywne. W odniesieniu do modeli bayesowskich oznacza to, że należy odrzucić przekonanie o ich normatywnym charakterze. Dyskutując z tym poglądem, argumentowałem za tym, że podział na deskryptywizm bayesowski

i preskryptywizm bayesowski jest pozorny i że ostatecznie mamy do czynienia z dwoma formami preskryptywizmu, tj. słabą i mocną. Wersja mocna preskryptywizmu zakłada postawę realistyczną w odniesieniu do bayesowskiego PP, zaś słaba implikuje postawę antyrealistyczną. Stwierdziłem, że uzasadnienie realizmu PP jest uzależnione od tego, czy da się wykazać, że mechanizmy poznawcze faktycznie mają charakter bayesowski. Argumentowałem za tym, że słaba wersja preskryptywizmu (preskryptywizm epistemiczny) jest zasadna w odniesieniu do PP. Zakłada ona możliwość wyjaśniania zjawisk przy użyciu twierdzenia Bayesa, nie przesądza jednak, czy zjawiska te są bayesowskie z natury. Oznacza to, że należy również uzasadnić realizm w odniesieniu do bayesowskiego PP, co w praktyce zakłada przyjęcie mocnego (ontycznego) preskryptywizmu. Zgodnie z nim, zjawiska wyjaśnia się za pomocą twierdzenia Bayesa, ponieważ są one bayesowskie. Jeżeli są one bayesowskie z natury, to znaczy, że *powinno* się je wyjaśniać przy użyciu modelowania Bayesa. Uzasadnienie tej tezy jest zadaniem kolejnych rozdziałów.

Rozdział III
Zasada energii swobodnej w przetwarzaniu predykcyjnym

Wstęp

W rozdziale tym podejmuję problem relacji pomiędzy PP a proponowaną przez Fristona FEP. Zasada ta zgodnie z zamysłem jej twórcy stanowi istotną i konieczną ramę eksplanacyjną dla wyjaśnienia dynamiki i złożoności organizmów żywych oraz realizowanych przez nie procesów poznawczych. Zwolennicy tej zasady twierdzą, że PP jest szczególnym przypadkiem realizacji mechanizmów minimalizacji energii swobodnej. Przeanalizuję argumenty zwolenników oraz przeciwników tego ujęcia. W tym celu odwołam się do sporu dotyczącego możliwości wykazania ciągłości pomiędzy procesami konstytuującymi życie a procesami konstytuującymi poznanie. Spór ten jest o tyle istotny, że stanowi swoiste „unaocznienie" możliwości eksplanacyjnych zasady energii swobodnej. W myśl stanowiska, które określę jako „nonkognitywistyczne", istnieje ciągłość na linii życie – umysł, którą można uzasadnić poprzez odwołanie się do pojęcia informacji syntaktycznej Shannona. Przeciwnicy tej tezy, których zaliczę do tzw. kognitywizmu, twierdzą, że wyjaśnienie poznania wymaga użycia innych narzędzi niż te, których używa się przy wyjaśnieniu życia, ponieważ, po pierwsze pojęcie informacji syntaktycznej nie wyczerpuje złożoności tych zjawisk, po drugie zaś stanowisko nonkognitywistyczne rodzi wiele problemów i niejasności. Zdaniem kognitywistów przy wyjaśnianiu życia, jak i umysłu należy odwołać się do pojęcia informacji semantycznej, które odrzucają nonkognitywiści. Podam wiele argumentów na rzecz tezy kognitywistycznej. Omawiając powyższy spór, wskażę na rolę, jaką zarówno kognitywizm, jak i nonkognitywizm przypisują FEP. Na podstawie tej analizy, a także dodatkowych argumentów, dokonam rewizji pretensji eksplanacyjnych wiązanych z tą zasadą i zastanowię się, na ile jest ona konieczna w wyjaśnieniu natury mechanizmów predykcyjnych. Ostatecznie odrzucę eksplanacyjne możliwości przypisywane FEP i będę bronił tezy, według której FEP stanowi istotną oraz niebanalną heurystykę badawczą dla analiz z zakresu PP, niemniej samo PP nie może zostać wyjaśnione w terminach proponowanych przez Fristona i jego zwolenników.

3.1. Uwagi wprowadzające

Można powiedzieć, że PP jest kombinacją dwóch idei (Gładziejewski, 2019; Klein, 2018): pierwsza z nich wyraża przekonanie, że koncepcja ta jest zbiorem różnych twierdzeń dotyczących tego, czym jest i jak funkcjonuje mózg (por. Clark, 2013b, 2016; Hohwy, 2013, 2015; Wiese, Metzinger, 2017); drugą zaś teza, zgodnie z którą u podstaw funkcjonowania mózgu czy umysłu, a właściwie całego ludzkiego organizmu, leży formalna koncepcja z biologii teoretycznej określana mianem „zasady energii swobodnej" (por. Friston, 2009, 2010, 2012; Friston, Kilner, Harrison, 2006; Friston, Stephan, 2007; Friston, Kiebel, 2009). Wśród badaczy nie ma jasności co do tego, w jakiej relacji pozostają względem siebie te dwie idee – nazwijmy je kolejno „architektoniczną" i „homeostatyczną", a także czy jest tak, że ujęcie architektoniczne musi być z konieczności rozpatrywane w kontekście ujęcia homeostatycznego. Odpowiedź na postawione pytania ściśle wiąże się z wyjaśnieniem tego, na czym polega ujęcie homeostatyczne i znajdująca się u jego podstaw FEP. Zanim jednak przejdę do charakterystyki FEP, należy powiedzieć kilka słów na temat problemu, który się tutaj pojawia.

Przypomnijmy: zgodnie z bronionym przez zwolenników PP przekonaniem podstawowym zadaniem mózgu (rozumianym jako hierarchiczny, wielopoziomowy bayesowski model generatywny) jest minimalizacja potencjalnych błędów predykcyjnych, które wynikają z różnicy pomiędzy naszymi oczekiwaniami dotyczącymi tego, jaki świat jest, a tym, jaki świat faktycznie jest. Charakterystyczna dla tego sposobu myślenia jest teza, według której percepcja jest zawsze i z konieczności ukierunkowana na realizację określonych działań danego organizmu w zmiennym, dynamicznym i niepewnym środowisku. Skoro percepcja jest zorientowana na działania, to zadaniem, przed jakim staje badacz, jest wyjaśnienie owego „zorientowania". Wyjaśnienie to polegać ma na tym, że wskazuje się na kolejne poziomy modelu, który generuje predykcje minimalizujące pojawiające się błędy predykcyjne na poziomach niższych. W myśl tej idei całe ludzkie poznanie można wyjaśnić, wykorzystując ten schemat: poczynając od pobudzenia komórek nerwowych, przez analizę kory mózgowej, motorykę, kontrolę ruchu, czynności odruchowe, aż po tzw. wyższe poziomy poznania, tj. myślenie, planowanie, podejmowanie decyzji itd. Trzeba stwierdzić, że ambicje eksplanacyjne, które niektórzy przypisują temu podejściu, są wielkie. Naturalne wydaje się zatem doszukiwanie się jakieś fundamentalnej biologicznej zasady, która pozwoli wyjaśnić predykcyjną naturę ludzkiego mózgu. Skoro bowiem ludzkie poznanie można wyjaśnić za pomocą bayesowskiego PP (por. Clark, 2013b; Hohwy, 2013, 2015a), to unifikacyjne ramy wyznaczane przez to podejście powinny dać się ująć, jak twierdzą niektórzy badacze, w jedną formalną regułę, która jest implementowana bądź realizowana na kolejnych poziomach modelu generatywnego. Ramą tą ma być, wyrażająca home-

ostatyczną naturę organizmów żywych, FEP. Zatem jeżeli ujęcie architektoniczne ma określoną moc eksplanacyjną, to pełnego jej uzasadnienia powinniśmy szukać w biologicznej, a jeszcze lepiej – metafizycznej naturze wyjaśnianych przedmiotów. Idąc tym tropem, należałoby stwierdzić, że eksplanacyjna moc podejścia architektonicznego wywodzi się z eksplanacyjnej i unifikującej mocy FEP.

Analizując powyższe uwagi, nasuwają się dwa wnioski: (1) relacja pomiędzy PP a FEP zależna jest od tego, jakie możliwości eksplanacyjne przypiszemy ujęciu architektonicznemu; oraz (2) określenie zależności ujęcia architektonicznego od homeostatycznego zakłada wiele rozwiązań dotyczących tego, co FEP wyjaśnia i w jaki sposób. Analizy przeprowadzone w niniejszym rozdziale będą miały na celu odpowiedź na dwie wymienione kwestie.

3.2. Pojęcie zasady energii swobodnej

Pojęcie energii swobodnej wywodzi się z termodynamiki i zasadniczo oznacza taką część energii całkowitej układu, jaka może zostać zamieniona na pracę w procesie o stałej temperaturze i stałej objętości. Jej zmiennymi są mierzalne temperatura i objętość. Kiedy oba te parametry są stałe, to energia osiąga minimum w stanie równowagi. Zmiana energii swobodnej jest równa maksymalnej pracy, jaką dany układ może wykonać w stałej temperaturze. Zdaniem Fristona, jego uczniów oraz współpracowników pojęcie to może być zasadnie używane w odniesieniu do żywych organizmów, czyli układów otwartych, tj. takich, jakie wchodzą w różne interakcje z otoczeniem, w którym żyją. Friston powiada, że wiele pozornie różnych aspektów struktury i funkcjonowania mózgu można rozumieć w kategoriach jednej, uniwersalnej i prostej zasady, mianowicie: minimalizacji energii swobodnej, która odzwierciedla (*reflects*) prawdopodobieństwo tego, co się dzieje na wejściach zmysłowych w odniesieniu do obecnego stanu mózgu (Friston, Stephan, 2007, s. 418). W ujęciu termodynamicznym energia swobodna jest miarą energii potrzebnej do wykonania określonej pracy. Zaś w ujęciu teoretyczno-informacyjnym, które jest charakterystyczne dla propozycji Fristona, energia swobodna odnosi się do różnicy pomiędzy tym, w jaki sposób świat jest modelowany, a tym, jaki faktycznie jest (Clark, 2016, s. 305). Zdaniem Fristona FEP odnosi się do wszystkich biologicznych systemów i wiąże z ich wrodzoną umiejętnością do przeciwstawiania się naturalnej tendencji do nieporządku. Systemy tego rodzaju są samoorganizujące się, a ich relacja ze środowiskiem ma charakter przyczynowości kołowej (*circular causality*). Ten typ przyczynowości zakłada istnienie różnych, dynamicznie zmieniających się stanów wewnętrznych i zewnętrznych. Stany zewnętrzne wobec systemu podlegają różnym zmianom, zaś stany wewnętrzne nie. Na drodze wymiany energii i materii z otoczeniem są one w stanie zachować integralność i porządek. Odbywa się

to jednak kosztem zwiększenia entropii, czyli miary stopnia nieuporządkowania otoczenia. W tym sensie wchodzi się tutaj w sferę życia oraz adaptacji systemów żywych, które aktywnie poszukują i wprowadzają warunki niezbędne dla własnego przetrwania.

Z punktu widzenia tak rozumianych systemów biologicznych korzystne jest zminimalizowanie energii swobodnej w celu utrzymania homeostazy rozumianej jako równowaga energetyczna (w sensie informacyjno-teoretycznym, a nie termodynamicznym) ze środowiskiem[1]. Minimalizacja energii swobodnej pozwala utrzymać żywy organizm w stanie względnie niskiej entropii. Skoro organizm pobiera informacje ze środowiska, to informacje te powinny być jak najmniej zaskakujące (*surprisal*), czyli stopień ich prawdopodobieństwa powinien być możliwie wysoki. Zatem każdy system biologiczny musi minimalizować stopień niepewności informacji docierających przez wejścia zmysłowe. Swobodna energia wiąże się z długoterminowym średnim zaskoczeniem (*long-term average 'surprisal'*), które odnosi się do ujemnego logarytmu prawdopodobieństwa pojawienia się określonych stanów zmysłowych opartych na takim, a nie innym modelu. Innymi słowy, „zaskoczenie" mówi o tym, jak mało prawdopodobne jest, aby pojawiło się dane zdarzenie w odniesieniu do takiego, a nie innego modelu świata. Przykładowo organizm zostaje zaskoczony wydarzeniem, którego zupełnie się nie spodziewał. W efekcie ma do czynienia z dużym błędem predykcyjnym. Zaskoczenie w perspektywie homeostatycznej jest niepożądane, zatem organizm zrobi wszystko, aby go uniknąć. Można powiedzieć, że zaskoczenie jest ujemnym logarytmem prawdopodobieństwa danego zdarzenia. W takim ujęciu entropia (w sensie teoretyczno-informacyjnym) jest długoterminowym średnim zaskoczeniem[2]. Minimalizacja energii swobodnej zwiększa wydajność modelu, w efekcie zmniejszając (długoterminowe średnie) zaskoczenie. Można powiedzieć, że w ten sposób organizmy opierają się drugiej zasadzie termodynamiki (Friston, Stephan, 2007; Ramstead, Badcock, Friston, 2017, 2018).

3.2.1. Pojęcie informacji Shannona

Pojęcie niepewności i zaskoczenia ściśle wiąże się z teorią informacji Shannona oraz nawiązuje do drugiej zasady termodynamiki i pojęcia entropii[3]. W 1948 roku

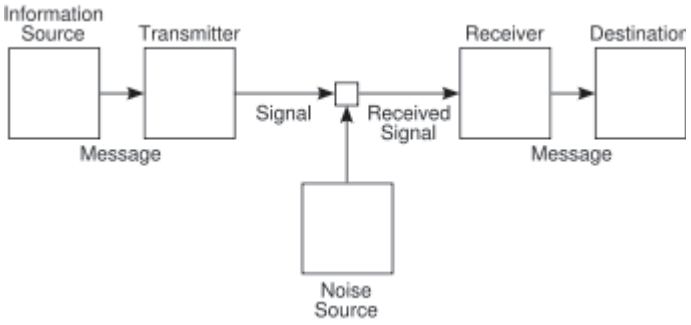
[1] Pojęcie homeostazy wywodzi się z prac francuskiego fizjologa Claude'a Bernarda, który stwierdził, że „niezmienność środowiska wewnętrznego jest warunkiem wolnego i samodzielnego życia" (1865).

[2] Zakładając, że proces, o którym mowa jest ergodyczny.

Ujęcie Fristona wywodzi się z rozumienia entropii w teorii informacji Shannona. Tam entropia oznacza średnią ważoną ilości informacji niesionej przez pojedynczą wiadomość. Wagami są prawdopodobieństwa nadania poszczególnych wiadomości.

[3] Według drugiej zasady termodynamiki entropia, czyli nieuporządkowanie układu, rośnie z czasem. Co prawda, druga zasada odnosi się do układów izolowanych bądź zamkniętych, ale w tym ujęciu chodzi przede wszystkim o to, że entropia zwiększa się wraz z przekazywaniem energii (Clark, 2017a, s. 3).

Claude E. Shannon opublikował artykuł pt. *Mathematical theory of communication*, w którym twierdził, że jeżeli chcemy wyjaśnić sposób, w jaki jest przesyłana informacja za pomocą środków komunikacji, należy opisać pewien idealny system komunikacji, w którym nie jest istotne znaczenie komunikatu, tylko jego struktura. Można ją przedstawić za pomocą prostego diagramu:



Rysunek 5. Struktura przetwarzania komunikatu (Shannon, 1948, s. 381)

W systemie komunikacji powinniśmy wyróżnić następujące elementy:

- źródło informacji (*information source*) – element, który wytwarza komunikat lub ich sekwencję;
- nadawca wiadomości (*transmitter*) – nadajnik, który w jakiś sposób przetwarza komunikat, wytwarzając sygnał odpowiedni do transmisji określony przez kanał;
- kanał (*channel*) – medium, które przesyła sygnał z nadajnika do odbiornika;
- odbiornik (*receiver*) – przetwarza sygnał na wiadomość, wykonując odwrotną funkcję wykonywaną przez wiadomość, tym samym dekodując ją;
- odbiorca (*destination*) – odbiera wiadomość.

Shannon analizuje system komunikacji w dwóch wersjach. W pierwszej, idealnej wersji zostaje pominięty problem szumów (*noise*), które są generowane przez jakieś zewnętrzne względem komunikatu źródła (*noise source*). W drugiej zaś zostaje wprowadzone pojęcie szumu i analizowany jest jego wpływ na możliwości przesyłania komunikatu. Shannon proponuje, aby kanał komunikatu opisać za pomocą pojęcia pojemności. Pojemność oznacza maksymalną miarę szybkości nadania komunikatu przez kanał[4]. Shannon twierdził, że miarą ilości danej informacji w komunikacie jest redukcja stopnia niepewności jego zajścia. Chodzi bowiem o to, w jaki sposób można zmierzyć niepewność jednego ze zdarzeń zawartych w zbiorze możliwych zdarzeń w sytuacji, w której dla każdego zdarzenia znane jest prawdopodobieństwo jego wystąpienia? Shannon proponuje,

[4] Jednostką pojemności jest bit na sekundę.

aby mierzyć informację za pomocą entropii, która ma być proporcjonalna do ilości informacji produkowanej przez źródło. Entropia zatem opisuje średni poziom niesionej informacji przez ciąg symboli. Na przykład: jeżeli miara entropii jest równa zeru, to wszystkie zdarzenia są niemożliwe z wyjątkiem jednego. Zatem im niższa jest entropia układu, tym wyższa jest jego przewidywalność, czyli pewność; zaś wraz ze wzrostem entropii zwiększa się poziom jego nieprzewidywalności, czyli niepewności (Shannon, Weaver, 1964, s. 13). Ilość informacji, jaką można wiązać z danym komunikatem, zależy zatem od jego nieprzewidywalności. Można ją określić w rachunku, jeśli porówna się prawdopodobieństwo komunikatu na wejściu z jego prawdopodobieństwem w miejscu jego przeznaczenia. Ilość informacji jest zatem miarą prawdopodobieństwa komunikatu. Informacja w takim ujęciu jest statystyczną własnością warunków generowania i transmisji sygnałów lub – inaczej mówiąc – jest to stochastyczna własność konkretnego zdarzenia – komunikatu. Wartość prawdopodobieństwa danego zdarzenia jest więc zależna od wartości prawdopodobieństwa zdarzenia go poprzedzającego[5].

Shannonowskie rozumienie informacji zostało wykorzystane przez Fristona do opisu pracy mózgu w języku termodynamiki (por. m.in. Friston, Stephan, 2007; Friston, 2009, 2010, 2012; Friston, Thornton, Clark, 2012). Rozważania Fristona opierają się na założeniach, według których (1) neuronauka może sformalizować relacje zachodzące pomiędzy informacjami przetwarzanymi przez mózg a energią; (2) pomiędzy termodynamiką i teorią informacji istnieje głęboka zależność przedmiotowa (por. Collell, Fauquet, 2015).

3.2.2. Homeostaza w zasadzie energii swobodnej

Friston formalizuje fakt samoorganizacji organizmów żywych w odniesieniu do informacji Shannona. Chodzi o to, że ze wszystkich możliwych stanów, w których organizm może się znaleźć, musi on znaleźć taki odpowiedni podzbiór stanów, jaki pozwoli mu przetrwać i skutecznie dokonywać wymiany energetycznej z otoczeniem. Istnieje zatem określony rozkład prawdopodobieństwa dla wszystkich możliwych stanów, w których organizm może się znaleźć. W dowolnym momencie rozkład ten (określony przez fenotyp danego organizmu) jest gwałtownie osiągany wokół pewnych wartości określających warunki organizmu niezbędne dla jego funkcjonowania i przeżycia. Na przykład ludzkie ciało ma wysokie prawdopodobieństwo osiągnięcia temperatury około 37°C, zaś niskie prawdopodobieństwo osiągnięcia temperatury 10°C. Matematycznie oznacza to, że rozkład prawdopodobieństwa

[5] Claude Shannon i Warren Weaver inspirowali się w swoich analizach opracowaną przez Markowa i Kołmogorowa teorią procesów stochastycznych i tzw. łańcuchów Markowa, która wiąże się z ideą, że prawdopodobieństwo danego zdarzenia jest wyznaczone tylko przez prawdopodobieństwo zdarzenia je poprzedzającego.

zmiennej temperatury ciała ma niską entropię w sensie Shannona. Innymi słowy, zdarzenie, w którym temperatura ciała wynosi około 37°C, jest mało zaskakujące dla organizmu, zaś sytuacja, w której osiąga ono 10°C, jest dla niego bardzo zaskakująca[6] (Bruineberg, Kiverstein, Rietveld, 2016, s. 5).

Zależność procesów termodynamicznych od homeostatycznych Friston i Stephan ilustrują przykładem płatków śniegu (2007). Porównują ze sobą normalne płatki śniegu z fikcyjnymi, uskrzydlonymi płatkami śniegu, które umieją się zaadaptować, oddziałując na otoczenie. Normalne płatki śniegu są w sposób bierny przepychane przez siły środowiskowe, aż ich temperatura osiąga pewien próg, przy którym przechodząc przez przejścia fazowe, tracą integralność i przekształcają się w krople wody. Inaczej jest w przypadku uskrzydlonych płatków śniegu. Te potrafią latać i utrzymywać się na niektórych wysokościach. Tym samym zachowują swoją temperaturę z dala od granicy topnienia. Utrzymanie odpowiedniej temperatury jest dla skrzydlatych płatków śniegu warunkiem przetrwania. Pomimo zmian w otoczeniu są one w stanie utrzymać w miarę stabilne środowisko na drodze wykonywania takich zachowań, jakie dostosowują je do nowych warunków. Zdaniem Fristona i Stephana można tutaj mówić o ich adaptacji, która oparta jest na dwóch warunkach: (1) termodynamiczny – takie zachowania w obrębie środowiska, jakie wykluczają fazę przechodnią i pozwalają utrzymać się z daleka od równowagi termodynamicznej; oraz (2) homeostatyczny – takie zachowania, jakie zmieniając relacje układu z otoczeniem, pozwalają zachować witalność fizjologiczną adekwatnie do zmieniających się warunków w określonych granicach[7]. Zdaniem cytowanych autorów podobne mechanizmy adaptacyjne można wiązać z rzeczywistymi organizmami, które wyewoluowały. Podobnie jak skrzydlate płatki śniegu unikają one przejść fazowych, aby utrzymać się z dala od równowagi termodynamicznej[8]. W tym celu służą im procesy minimalizacji energii swobodnej, czyli procesy homeostatyczne w rozumieniu Fristona. Należy powiedzieć, że FEP pozwala utrzymać homeostazę (rozumianą w terminach teoretyczno-informacyjnych) na odpowiednim poziomie, realizując mechanizm, który (1) wykrywa nawet niewielkie odchylenia od optimum; (2) prowadzi do takich reakcji, jakie pozwalają na wyrównanie tego odchylenia; oraz (3) umożliwia przerwanie tych reakcji wtedy, kiedy organizm powraca do optymalnego poziomu homeostazy.

Wróćmy teraz na moment do pojęcia zaskoczenia. Należy powiedzieć, że stopień zaskoczenia daną informacją (*surprisal of information*) jest mierzony stopniem jej prawdopodobieństwa. Z tego powodu ideę samoorganizujących się minimalizują-

[6] Oczywiście to, czy taka, a nie inna temperatura ciała jest zaskakująca dla danego organizmu, zależy od określonego gatunku, czyli jego genotypu i fenotypu.

[7] Jeden z twórców cybernetyki Kenneth Craik podkreślał, że podstawowe zadanie samoregulacji wymaga, aby systemy biologiczne koordynowały swoje zachowanie z tymi cechami środowiska zewnętrznego, które mogą je osłabiać (Craik, 1960). Na temat związku analiz Fristona z cybernetyką zob. rozdział *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego*.

[8] Jako przykład Friston i Stephan podają wybór chemotaksji w organizmach jednokomórkowych.

cych energię swobodną systemów biologicznych Friston wiąże z modelowaniem bayesowskim (Friston, 2010, 2012). W takiej perspektywie modele mające minimalną ilość energii swobodnej dostarczają dokładne wyjaśnienie danych przy określonych kosztach, które są tutaj rozumiane jako rozbieżność pomiędzy gęstością wariacyjną a wcześniejszymi przekonaniami dotyczącymi ukrytych stanów środowiska. Jak zatem rozumieć minimalizację energii swobodnej w kategoriach bayesowskiej minimalizacji błędów predykcyjnych?

Minimalizacja energii swobodnej polega na takiej zmianie wewnętrznych parametrów modelu, aby oszacować możliwe prawdopodobieństwo przyczyn pobudzeń zmysłowych (*causes of sensation*) (Friston i in., 2010). Innymi słowy, energia swobodna (w sensie teoretyczno-informacyjnym) zostaje zminimalizowana wtedy, gdy zostaną zmienione predykcje dotyczące źródeł informacji statystycznej docierającej przez wejścia zmysłowe. Można tego dokonać albo (1) przez zmianę własności modelu – zmiana przyjętych predykcji; takie działanie Friston określa jako „wnioskowanie percepcyjne" (*perceptual inference*); albo (2) przez zmianę w środowisku – „wnioskowanie aktywne" (*active inference*), czyli działanie, które ingeruje w przyczynową strukturę świata (Friston, 2010, s. 129). Percepcja redukuje energię swobodną, zmieniając predykcje lub wybrane parametry modelu, działanie zaś dokonuje takiej minimalizacji na drodze zmiany informacji docierającej do modelu[9]. Z tego powodu opisywane przez Fristona systemy biologiczne należy interpretować w kategoriach aktywnych podmiotów (*active agents*)[10] (Friston, 2013), które minimalizują energię swobodną, aby utrzymać swoje procesy życiowe na optymalnym poziomie. Aktywny podmiot minimalizuje błąd predykcyjny za pomocą probabilistycznego modelu generatywnego. Minimalizacja błędu predykcyjnego spełnia zatem określone funkcje normatywne w odniesieniu do podmiotu. Po pierwsze jest ona konieczna dla utrzymania równowagi homeostatycznej; po drugie zobowiązuje (*obliges*) podmiot do utworzenia takich predykcji dotyczących stanów świata, jakie pozwolą na wybór odpowiednich działań na drodze optymalizacji statystycznej informacji pochodzącej z wejść zmysłowych (Friston i in., 2010, s. 233). Można zatem stwierdzić, że predykcje minimalizujące błąd predykcyjny nie są prostymi, odgórnie generowanymi komendami motorycznymi (*top-down motor commands*), ale uprzednimi oczekiwaniami dotyczącymi informacji percepcyjnej i priopercepcyjnej. Oznacza to, że „ośrodkowy system nerwowy nie jest podzielony na systemy motoryczny i sensoryczny, ale że jest on maszyną inferencyjną, która

[9] Wnioskowaniem aktywnym szczegółowo zajmuję się w rozdziale *Mechanizmy normatywne a działania w przetwarzaniu predykcyjnym*.

[10] Pojęcie agenta byłoby tutaj bardziej uzasadnione ze względu na skojarzenia semantyczne wskazujące na jego działania i aktywny charakter. Pojęcie to jednak w polskiej literaturze filozoficznej nie jest popularne i budzi skojarzenia z innymi dyscyplinami, np. socjologią. Z tego powodu korzystam z tradycyjnego, ale dużo bardziej obciążanego semantycznie i problemowo, pojęcia podmiotu.

dostarcza predykcji dotyczących optymalnych działań w terminach oczekiwanych konsekwencji" (Friston i in., 2010, s. 235).

Po tych wstępnych uwagach przejdę teraz do tzw. łańcuchów Markowa, które część badaczy łączy z FEP, wykazując tym samym ścisłą zależność ujęcia homeostatycznego z architektonicznym, czyli z PP.

3.3. Koce Markowa a przetwarzanie predykcyjne

Zdaniem niektórych autorów (por. m.in. Bruineberg, Rietveld, 2014; Bruineberg, Kieverstein, Rietveld, 2016; Friston, Kiebel, 2009; Hohwy, 2015a, 2015b, 2016, 2017; Kirchhoff i in., 2018; Ramstead, Kirchhoff, Friston, 2019) FEP wyznacza istotną ramę eksplanacyjną, która pozwala z jednej strony ugruntować PP na biologicznych podstawach, z drugiej zaś lepiej wyjaśnić jego naturę oraz zastosowania. Przykładowo Hohwy wprost stwierdza, że FEP koresponduje z architekturą modeli generatywnych rozwijanych w ramach PP, oraz, że tym samym jest mówienie o minimalizacji błędów predykcyjnych i minimalizacji energii swobodnej (Hohwy, 2015a, s. 1, 8, 20–21). Zdaniem Setha (2015, s. 5) PP można traktować jako specjalny przykład zastosowania FEP; zaś Kirchhoff twierdzi, że o PP należy myśleć w terminach minimalizacji energii swobodnej (Kirchhoff i in., 2018, s. 3). Powiązanie obu tych perspektyw ma być obiecujące, ponieważ, jak argumentują wspomniani autorzy, pozwoli ono m.in. na pełniejsze wyjaśnienie natury umysłu (Hohwy, 2015a), wewnętrznej dynamiki mózgu (Friston, Kiebel, 2009) oraz złożonych relacji na linii podmiot (organizm) – środowisko (Bruineberg, Rietveld, 2014). Możliwości powiązania perspektyw homeostatycznej i architektonicznej (predykcyjnej) upatruje się w wykorzystaniu tzw. koców Markowa[11].

W myśl omawianych wykładni należy powiedzieć, że stany wewnętrzne danego organizmu można zdefiniować jako koce Markowa, które oddzielają system od jego środowiska. Stany te są zatem niezależne (w sensie statystycznym) od stanów otoczenia. Oznacza to, że stany wewnętrzne mają za zadanie minimalizować energię swobodną, aby zachować strukturalną i funkcjonalną integralność, która prowadzi do homeostazy i samoorganizacji. Warunkiem minimalizacji energii swobodnej jest zatem ścisła separacja wnętrza systemu od jego otoczenia, czyli zdolność tego systemu do odróżniania siebie od środowiska (Friston, 2013a)[12]. Koce Markowa umożliwiają wyznaczenie granic systemu.

[11] Pojęcie to wprowadziłem formalnie w podrozdziale *Podstawowe pojęcia*; por. także podrozdział *Internalizm a koce Markowa*.

[12] Istnienie koców Markowa nie pozwala na przekazywanie dowolnej ilości informacji do wnętrza systemu. Tym samym są one gwarantem utrzymania możliwie niskiego poziomu entropii (por. Ashby, 1963, s. 247).

Hohwy (2014, s. 18) powiada, broniąc swojej wersji PP, że:

> umysł rozpoczyna się tam, gdzie dostarczane są dane zmysłowe przez receptory eksteroceptywne[13], proprioceptywne[14] i interoceptywne[15], a kończy tam, gdzie trafiają predykcje propioceptywne, czyli głównie w rdzeniu kręgowym.

Tak rozumiany model generatywny jest samouwierzytelniający się (*self-evidencing*), co oznacza, że formułowane na jego gruncie hipotezy (oparte na wnioskowaniu do najlepszego wyjaśnienia) najlepiej tłumaczą pewne informacje, jeżeli traktuje się je jako dowody na rzecz twierdzeń, które dane hipotezy głoszą (por. Hempel, 1965, s. 372–374; Hohwy, 2014, s. 5). Przykładowo: kiedy wracałem do domu, zobaczyłem, że mam wybitą szybę w oknie. W tym przypadku muszę postawić jakąś hipotezę, aby jak najlepiej zminimalizować błąd predykcyjny, jakim jest informacja o wybitej szybie. Nie znam przyczyn, które doprowadziły do tej sytuacji, więc muszę założyć kilka prawdopodobnych scenariuszy, w tym np. ten, który zakłada włamanie do mojego domu. Stawiam zatem hipotezę, według której szybę zbił złodziej. W tej sytuacji zbita szyba stanowi część dowodu na rzecz hipotezy o włamaniu. Na podobnej zasadzie działa model generatywny zdaniem Hohwy'ego. Informacje docierające do modelu mają jedynie wartość jako siła dowodowa w uzasadnianiu predykcji, które model generuje. Bez istnienia koców Markowa, które można wiązać z kolejnymi poziomami modelu generatywnego, nie byłoby możliwe uzasadnianie predykcji minimalizujących potencjalne błędy predykcyjne. Clark (2017a, s. 4), komentując koncepcję Hohwy'ego, zwraca uwagę na dwie rzeczy: (1) organizmy żywe traktowane są jako samouzasadniające się systemy minimalizacji energii swobodnej; oraz (2) ludzki mózg jest w istocie implementacją procesu minimalizacji błędów predykcyjnych. Takie ujęcie pozwala, jak sądzę, mówić o mocnym internalizmie w kwestii rozumienia ludzkiego poznania przez Hohwy'ego oraz o sprowadzaniu całej aktywności ludzkiego mózgu i umysłu do minimalizowania niepewności i zaskoczenia (por. podrozdział *Internalizm a koce Markowa*).

Kirchhoff i współautorzy (Kirchhoff, 2018; Kirchhoff i in., 2018) proponują zupełnie inne wykorzystanie koców Markowa w PP i FEP. Zwracają uwagę na to, że statystycznie rzecz ujmując, istnienie koców Markowa pozwala mówić jedynie o warunkowym wpływie stanów otoczenia na stany wewnętrzne organizmu[16]. Ten warunkowy wpływ należy wiązać z wnioskowaniem aktywnym, które w tym kontekście można rozumieć jako tendencję losowych układów dynamicznych do

[13] Ekstrocepcja – czucie powierzchowne (temperatury, bólu, dotyku itd.).
[14] Propriocepcja – zmysł kinestetyczny, czucie głębokie (dostarczanie informacji o pozycji kończyn, głowy itd.).
[15] Introcepcja – czucie trzewne (odbierane w ścianach naczyń krwionośnych i trzewiach).
[16] Chodzi o to, że w celu opisania natury systemu autopojetycznego, który jest nieustannie sprzężony ze swoim otoczeniem, nie potrzeba odwoływać się do zmian obecnych w środowisku (por. Chemero, 2012, s. 53–54).

minimalizacji ich energii swobodnej. Oznacza to, że procesy samoorganizacji w kocach Markowa oparte są na samodowodzącej się dynamice, która leży u podstaw życia, czyli minimalizacji energii swobodnej. Redukcja energii jest zatem tym samym, co optymalizacja dowodów dla danego modelu czy systemu. W ten sposób relację pomiędzy stanami wewnętrznymi a zewnętrznymi organizmu (przy użyciu koców Markowa), opisywalną w terminach minimalizacji energii swobodnej, da się powiązać z bayesowskim PP. W myśl tej interpretacji koce Markowa znajdują się na każdym poziomie modelu generatywnego, a minimalizację błędów predykcyjnych (energii swobodnej), czyli różnicę pomiędzy predykcjami a pobudzeniami na wejściach zmysłowych, można tłumaczyć wewnętrznymi stanami organizmu (Kirchhoff i in., 2018, s. 3–4; por. także Ramstead, Badcock, Friston, 2018). Kluczowa dla tego ujęcia jest kwestia wnioskowania aktywnego, czyli takiej ingerencji w przyczynową strukturę świata, jaka pozwala na uprawdopodobnienie przyjętych hipotez predykcyjnych. Dzieje się to na drodze minimalizacji zaskoczenia, czyli innymi słowy, zwiększenia mocy dowodowej modelu. Nacisk na konieczność uwzględniania w analizie wnioskowania aktywnego przypomina o tym, że przeprowadzenie wnioskowania Bayesa nie opiera się tylko na (wewnętrznej) aktywności neuronalnej, ale także na dynamicznej interakcji ze środowiskiem[17]. Można zatem mówić o ucieleśnieniu modelu generatywnego, polegającym na selektywnym testowaniu (*sampling*) danych zmysłowych w celu minimalizowania zaskoczenia związanego z nieznajomością ich ukrytych przyczyn.

Prace Kirchhoffa, o czym wspomniałem w rozdziale *Pojęcie przetwarzania predykcyjnego*, należy wiązać z radykalnym, inspirowanym enaktywizmem, podejściem do PP. Powiada on, że organizmy żywe, jako posiadające koce Markowa, nieustannie odtwarzają autopojetyczne procesy samoorganizacji, które dopóki dany organizm żyje, dopóty wymuszają różnicę pomiędzy funkcjonującym systemem a jego otoczeniem (Kirchhoff i in., 2018, s. 7). Procesy te są zatem procesami konstytuującymi identyczność danego organizmu. Wykorzystanie koców Markowa w wyjaśnieniu funkcjonowania samoorganizujących się i samowytwarzających się systemów pozwala na nowo odczytać analizę żywej komórki przeprowadzoną przez Francisca Varelę w artykule *Patterns of life: Intertwining identity and cognition* (1997). Powiada on, że każda bakteria za pośrednictwem półprzepuszczalnej błony jest oddzielona od swojego molekularnego otoczenia. Jest ona jednocześnie tworzona i utrzymywana przy życiu przez sieć procesów (absorbujących różne składniki odżywcze czy wydalających odpady), które wyznaczają granicę pomiędzy żywym organizmem a środowiskiem. Procesy te, wiązane przez Varelę z tzw. siecią metaboliczną (Varela, 1997, s. 75), można interpretować w kategoriach prostego

[17] W podrozdziale *Konserwatywne przetwarzanie predykcyjne* wykazuję, że stanowisko Hohwy'ego (tj. konserwatywne PP) zakłada konieczność odniesienia się do środowiska, niemniej inaczej to rozumiem niż zwolennicy radykalnego PP.

koca Markowa. Jest on jednocześnie odpowiedzialny (w sensie konstytutywnym) za autopojetyczny proces utrzymywania identyczności komórki samej z sobą. Identyczność autopojetyczna jest rozumiana procesualnie i wiąże się z operacyjnym zamknięciem (wyznaczanym przez procesy Markowa) na procesy obecne w środowisku, czyli te realizowane poza błoną komórkową. Nie należy jej w żaden sposób wiązać z jakimś opisem strukturalnym, ale z byciem nośnikiem określonych własności mentalnych bądź psychicznych. Dla żywego organizmu istotne jest tylko to, że jest on oddzielony od swojego środowiska ze względu na własną tożsamość, która definiuje dany organizm jako autopojetyczną jedność (Varela, 1997, s. 79–80). Kirchhoff, komentując ten fragment, stwierdza, że owa jedność jest ograniczeniem, które wymusza warunkową niezależność stanów wewnętrznych od zewnętrznych. Zaś ta niezależność jest warunkiem udanego sprzężenia pomiędzy systemem a jego otoczeniem. Sprzężenie to umożliwia organizmom adaptacyjne i aktywne działanie w środowisku, które prowadzi do zmiany charakteru relacji wiążącej system i otoczenie, w którym on żyje (Kirchhoff i in., 2018, s. 7).

3.4. Energia swobodna a ciągłość życia i umysłu

Zdaniem Fristona każdy żywy organizm jest procesem ergodycznym, czyli takim, dla jakiego statystyki uśredniane po pewnym czasie są takie same, jak jego statystyki uśredniane po realizacji procesu. Zdaniem Fristona posiadanie koca Markowa jest warunkiem bycia procesem ergodycznym (Friston, 2013c). Organizm jako dynamiczny system sprzężony ze środowiskiem musi w nim działać, realizując aktywne wnioskowanie, aby osiągnąć pożądany stan homeostazy, czyli możliwie dobrze minimalizować swoje zaskoczenie. Fristonowska FEP ma zapewnić formalne i teoretyczne ramy wyjaśniające konstytutywne sprzężenie mózgu i ciała ze środowiskiem, które jest warunkiem homeostazy (Allen, Friston, 2018, s. 2476). Eksplanacyjną wartością FEP zajmę się w dalszej części tego rozdziału, teraz natomiast przyjrzę się sposobowi, w jaki niektórzy badacze używają tej zasady, aby wykazać ciągłość zachodzącą pomiędzy biologicznym życiem i odpowiedzialnym za wyższe procesy poznawcze umysłem[18]. Jest to również pouczający przykład zastosowania FEP w określonym polu problemowym.

[18] W niniejszym rozdziale koncentruję się na dwóch skrajnych stanowiskach (kognitywizm i nonkognitywizm) w kwestii problemu ciągłości życia i umysłu oraz zagadnienia informacji semantycznej. Należy jednak podkreślić, że istnieje wiele różnych propozycji pośrednich, które w mniejszym lub większym stopniu zakładają istnienie takiej ciągłości bądź argumentują za jej brakiem (por. Allen, 2018; Allen, Friston, 2018; Korbak, 2019b; Villalobos, Dewhurst, 2018).

3.4.1. Ujęcie enaktywistyczne

Michael Kirchhoff i Tom Froese twierdzą, że fristonowską FEP należy wiązać z teorią autopojezy rozwijaną przez Varelę i Maturanę. Nie wdając się w analizy tego, jak rozumieć umysł (o czym powiem później), proponują, że należy odróżnić dwa główne poglądy dotyczące związków zachodzących pomiędzy życiem a umysłem (Kirchhoff, Froese, 2017; Kirchhoff, 2018). Pierwszy z nich określają jako kognitywizm. Zwolennicy tego ujęcia głoszą niezależność życia i umysłu. Łączą energię swobodną (w sensie teoretyczno-informacyjnym) z informacją semantyczną, ograniczając tym samym zakres organizmów, którym można przypisać proste funkcje poznawcze (*mentality*). Kognitywizmowi Kirchhoffa i Froese przeciwstawiają nonkognitywizm, który ma trzy odmiany: (1) termodynamiczną, łączącą FEP z fizyką statystyczną i stosowaną w wyjaśnieniu dynamiki samoorganizacji w organizmach, unikających równowagi termodynamicznej (por. Friston, 2009, 2010); (2) ewolucyjną, która opiera się na stwierdzeniu braku ciągłości pomiędzy systemami żyjącymi a systemami poznawczymi. W tym ujęciu procesy poznawcze wymagają istnienia złożonych mechanizmów neuronalnych, które nie występują u prostych organizmów. Oznacza to, że żywy organizm nie jest z konieczności systemem poznawczym. Przedstawicielem takiego stanowiska ma być, zdaniem Kirchhoffa i Froese, Clark (por. Clark, 2017a); oraz (3) koncepcję, określoną mianem „enaktywistycznej”[19], która podkreśla silną ciągłość zachodzącą pomiędzy życiem a umysłem. Z perspektywy prowadzonych tutaj analiz, istotne jest porównanie kognitywizmu z enaktywizmem.

Argumentacja zwolenników enaktywistycznego podejścia do FEP opiera się na dwóch przesłankach. Pierwsza z nich dotyczy natury umysłu. Druga zaś analizowania umysłu w kategoriach autopojezy i adaptacji. Zacznijmy od pojęcia umysłu. Kirchhoff i Froese wykorzystują zaproponowane przez Hutto i Myina pojęcie umysłu prostego (*basic mind*) (Hutto, Myin, 2013). Zdaniem tych autorów umysł w swoim najbardziej podstawowym wymiarze (tj. na poziomie wczesnej onto- i filogenezy) jest pozbawiony jakiejkolwiek treści. Można go wyjaśnić jedynie w kategoriach złożonych i dynamicznych interakcji sensomotorycznych ze środowiskiem, nie powołując się przy tym na żadne reprezentacje. Takie pojęcia jak „treść” czy „reprezentacja” nie mają wartości eksplanacyjnej. Treść w umyśle jest nabywana jedynie w kontekście praktyk społeczno-lingwistyczno-kulturowych. Czym zatem jest treść? Zdaniem Hutto i Myina treścią jest wszystko to, co ma określone warunki spełniania (*conditions of satisfaction*), czyli musi być albo prawdziwe, albo fałszywe, poprawne bądź nie. Tutaj pojawia się tzw. trudny problem treści (*hard problem of content*), który podważa możliwość naturalistycznego uzasadnienia zachodzenia relacji pomiędzy

[19] Będę ją odróżniał od ujęcia enaktywno-ekologicznego, którego przedstawicielami są Bruineberg, Kiverstein czy Rietveld. Krytyczną analizę ich prac przedstawiam w podrozdziale *Predykcje skierowane na afordancje*.

własnościami, które mają treść, a własnościami fizycznymi (Hutto, Myin, 2013, s. 69). Omawiani autorzy twierdzą, że nie ma takiego naturalistycznego pojęcia treści, jakie rozwiązywałoby trudny problem treści, ponieważ wszelkie zaproponowane w literaturze pojęcia informacji oparte są na kowariancji, czyli wielkości, która charakteryzuje wspólne zmiany dla dwóch zmiennych. Pojęcie informacji semantycznej z tego powodu również nie pozwala wyjaśnić sposobu, w jaki umysł zyskuje treść. Postulowanie istnienia takiego rodzaju informacji jest raczej założeniem niż wyjaśnieniem istnienia własności semantycznych albo zawierających treść (Hutto, Myin, 2013, s. 67). Z tego też powodu autorzy *Radicalizing enactivism: Basic minds without content* posługują się pojęciami prostego poznania, które nie ma charakteru fenomenalnego, tj. odnoszącego się do tych własności stanów mentalnych (jak wrażenia czy dane zmysłowe), dzięki którym są one „o czymś" i mają charakter przekonaniowy, ale ma charakter intencjonalny, czyli wykazuje celowe ukierunkowanie. Owa celowość jednak nie implikuje intencjonalności w sensie osobowym, tj. świadomego odnoszenia się do jakiejś przedmiotowości, ale raczej sposób, w jaki organizm aktywnie angażuje się w swoje środowisko. Można tutaj, jak sądzę, mówić o minimalnej intencjonalności biologicznej (Hutto, Myin, 2013, s. 30), ukonstytuowanej ewolucyjnie.

W tym kontekście nawet najbardziej prymitywne organizmy żywe, np. bakterie, można scharakteryzować jako mające proste umysły. Dowiedzeniu tego ma służyć uwaga Petera Godfreya-Smitha i Kima Sterelnego, że jedyną charakterystyką informacji biologicznej ma być ta, którą podał Shannon, a która pojawia się zawsze wtedy, gdy istnieje jakiś przypadek, relacja przyczynowo-skutkowa czy określona korelacja (Godfrey-Smith, Sterelny, 2016). Froese i Kirchhoff nie zwracają jednak uwagi na pełny kontekst tej wypowiedzi. Mianowicie: Godfrey-Smith i Sterelny stwierdzają, że informacja Shannona spełnia pewne minimalne warunki, jakie można nałożyć na jakąkolwiek informację. W tym sensie przykładowo dym niesie informację o pożarze. Oznacza to, że sygnał przenosi informację o źródle w tym sensie, że możemy przewidzieć stan źródła na podstawie sygnału. Nie powiemy jednak, że z tego powodu dym jest żywy. Shannon takie pojęcie informacji wiązał przede wszystkim z technologiami komunikacyjnymi. I to jest ważna wskazówka, bowiem jeżeli uznać wymagania nakładane na informację przez Shannona za kryteria informacji biologicznej, to zanika różnica pomiędzy organizmami żywymi, takimi jak bakterie, zwierzęta czy nawet rośliny, a sztucznym życiem, zjawiskami fizycznymi (jak dym) lub wszelkimi urządzeniami przetwarzającymi informację, tj. prędkościomierzami, termometrami czy kalkulatorami itd.[20]

[20] Podobny zarzut można postawić koncepcji prostych umysłów, które są na tyle szeroko ujęte, że prostą umysłowość (*mentality*) można również przypisać innym obiektom niż organizmy żyjące. Z innej strony Tomasz Korbak (2015) pokazał, że prostych umysłów nie ma, bo bakterie też muszą mieć z punktu widzenia Hutto i Myina język, czyli system spełniający założenia Skyrmsa (por. Skyrms, 2010). Wiąże się to ze słabością dystynkcji na umysły proste i umysły oparte na rusztowaniach społeczno-językowych (*scaffolded minds*).

Tych słabości omawiani autorzy zdają się nie spostrzegać i raczej sugerują, że z trudnym problemem treści dobrze sobie radzi pojęcie informacji wzajemnej (*mutual information*) (Kirchhoff, Froese, 2017, s. 7)[21]. Pojęcie informacji wzajemnej ma w sposób satysfakcjonujący ujmować intuicje zawarte we fristonowskiej FEP: wzajemność jest kowariancją w tym sensie, że dwie zmienne są ze sobą skorelowane w taki sposób, że informacja niesiona przez jedną redukuje poziom zaskoczenia u drugiej. Tak rozumiana informacja optymalizuje model. Nie potrzeba się tutaj odwoływać do żadnych stanów wewnętrznych organizmu, manipulacji reprezentacjami czy treści semantycznych[22]. Należy jednak zwrócić uwagę na fakt, że informacja wzajemna jest uogólnioną kowariancją w takim samym sensie jak informacja semantyczna. Niemniej od informacji semantycznej wymaga się spełnienia dodatkowych kryteriów.

Różnica pomiędzy umysłem a życiem, twierdzą Kirchhoff i Froese, jest jak różnica pomiędzy informacją semantyczną a niesemantyczną[23]. Ta pierwsza jest ukonstytuowana na drugiej za sprawą procesów zewnętrznych względem organizmu. Proste umysły są silnie powiązane z biologią danego organizmu. Są one intencjonalne i celowo skierowane na swoją niszę biologiczną, na którą reagują (por. Hutto, Myin, 2013, s. X). Owa reakcja nie jest jednak zależna od jakiejś interpretacji, rozumienia czy jakkolwiek rozumianej aktywności kierowanej reprezentacjami. Ową celową intencjonalność można wyjaśnić przez odniesienie do historii selekcji naturalnej, ponieważ jej jedynym zadaniem jest wzbudzenie określonych reakcji na takie, a nie inne rzeczy czy zdarzenia (Hutto, 2006). Z tego to powodu ktoś mógłby stwierdzić, że intencjonalne odniesienie czy ukierunkowanie ma wymiar normatywny (Kirchhoff, Froese, 2017, s. 13), tzn. że można mówić tutaj o poprawnym bądź niepoprawnym reagowaniu na środowisko. Przykładowo: organizm reaguje na zmianę temperatury termoregulacją, np. poci się, gdy temperatura wzrasta powyżej pewnej średniej, albo trzęsie, gdy temperatura otoczenia spada poniżej tej średniej. Zaburzenie mechanizmów termoregulacji można zatem traktować jako dysfunkcję[24]. Kirchhoff i Froese nie chcą jednak mówić o normatywności funkcji

[21] Informacja wzajemna jest miarą zależności pomiędzy dwiema zmiennymi losowymi. Inaczej mówiąc: informacja wzajemna mierzy, ile informacji o danej zmiennej można poznać, znając inną, czyli o ile poznanie jednej z tych zmiennych zmniejsza niepewność o drugiej. Jeśli zmienne są od siebie niezależne, to ich wzajemna informacja jest zerowa (znajomość jednej nie mówi niczego o drugiej). Jeśli są identyczne, to każda zawiera pełną wiedzę o drugiej.

[22] Interesujące w tym kontekście jest to, że podobieństwo (w reprezentacjach opartych na podobieństwie, np. S-reprezentacjach) można mierzyć w kategoriach informacji wzajemnej. Za tę uwagę dziękuję Marcinowi Miłkowskiemu.

[23] Według tego poglądu pierwotną i podstawową funkcją układu nerwowego jest utrzymanie łączności pomiędzy percepcją a działaniem. Wizja ta, pomimo swojej popularności i ogólnego uznania, nie jest jedyną. Drugi pogląd na podstawową funkcję działania układu nerwowego wiążę się z samoformowaniem działania, czyli pojawieniem się pewnej makroczynności i mnóstwa różnych mikroczynności (por. Keijzer, van Duijn, Lyon, 2013; Godfrey-Smith, 2018, s. 39–40). Zestawienie obu tych wizji jest istotne, ponieważ pokazuje, że wyjaśnianie jednym prostym schematem poznania może prowadzić nie tylko do uproszczeń, ale zafałszowania całości.

[24] Normatywnością funkcji zajmuję się w rozdziale *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego*.

albo mechanizmów. W zamian proponują ujęcie normatywności, które odnosi się do procesów samoorganizacji i autopojezy realizowanych przez organizm jako pewną inherentną całość. Tutaj ujawnia się różnica pomiędzy systemami żywymi a artefaktami, takimi jak termostat. Te ostatnie realizują funkcje normatywne za sprawą swoich projektantów. Żywe systemy autopojetyczne są autonomiczne w sensie energetycznym. Są zatem źródłem swojej własnej wewnętrznej normatywności, którą należy wiązać z utrzymywaniem stanów wewnętrznych przez ich własne koce Markowa, na drodze realizacji wnioskowania aktywnego. Życie jest procesem, dzięki któremu organizm unika stanu równowagi termodynamicznej, inwestując energię w minimalizowanie niepewności. W ten sposób definiuje on granicę pomiędzy tym, co należy do systemu, a tym, co należy do jego środowiska. Owa granica określa autonomię jednostki.

Powiązanie jednostki ze środowiskiem można opisać w kategoriach holistycznej normatywności. Bezpośrednie intencjonalne odniesienie opiera się na procesach autopojezy oraz adaptacji w środowisku opartej na FEP, które prowadzą do indywiduacji. Ciągłość życia i umysłu jest zatem wynikiem bezpośredniej normatywnej relacji zachodzącej pomiędzy podstawowymi wymiarami życia (autopojetycznymi) a podstawowymi wymiarami umysłowości (bezpośrednio intencjonalnej, pozbawionej treści i semantyki) (Kirchhoff, Froese, 2017, s. 16). Za pośrednictwem swojego koca Markowa system jest operacyjnie domknięty, tzn. jego dynamika, oddzielająca stany wewnętrzne i zewnętrzne, umożliwia jego jednostkowienie. Kirchhoff (2018), powołując się na analizy Ezequiela Di Paolo i Evana Thompsona (por. Thompson, 2011), dodaje, że sama rama wyznaczana przez teorię autopojetyczną nie pozwala w pełni wyjaśnić istnienia umysłów. Dopiero wzbogacenie jej przez FEP czyni zadość temu oczekiwaniu. Należy się zgodzić, powiada Kirchhoff, że autopojeza jest koniecznym i wystarczającym warunkiem dla życia. Z tego jednak nie wynika, że jest wystarczającym warunkiem dla pojawienia się umysłu. Dzięki temu, że w ramach FEP poznanie jest zawsze zorientowane na przyszłość, zyskuje ono charakter adaptacyjny: „systemy, które minimalizują swoją energię swobodną, odpowiadają na zmiany w środowisku w sposób adaptacyjny" (Friston, Stephan, 2007, s. 428). Znaczy to, że cel, jakim jest minimalizacja zaskoczenia i niepewności, wymaga regulacji, która opiera się na ciągłych przewidywaniach. Dzieje się to na drodze realizacji wnioskowania aktywnego, czyli takiego próbkowania środowiska, jakie ma charakter epistemiczny. Dobrym przykładem są eksploracyjne ruchy sakkadowe gałki ocznej. W ten sposób życie wiąże się z poznaniem: procesy homeostatyczne są współregulowane przez procesy epistemiczne. Zdaniem Kirchhoffa takie powiązanie FEP i teorii autopojezy pozwala w satysfakcjonujący sposób wykazać ciągłość zachodzącą pomiędzy życiem a umysłem. Pogląd ten można określić mianem „biopsychizmu" (Kirchhoff, 2018, s. 2531–2532).

3.4.2. Krytyka ujęcia enaktywistycznego

Koncepcji proponowanej przez Kirchhoffa i Froese można postawić wiele zarzutów. Pierwszy z nich dotyczy zastosowania takich pojęć jak życie czy prosty umysł oraz postulatu głoszącego konieczność powiązania ze sobą desygnatów tych pojęć. Trudno jednoznacznie scharakteryzować, co ci autorzy rozumieją przez życie. W ich tekstach próżno szukać innego określenia niż biologiczna samoorganizacja. Jest to jednak ujęcie czysto opisowe i nie można przypisać mu specjalnej mocy eksplanacyjnej. Co więcej, samo mówienie o życiu jako takim jest mocno problematyczne, ponieważ prowadzi do urzeczowienia procesów, które przebiegają w organizmach. „Nie istnieje ono [życie] jako niezależny byt. Naukowo można zajmować się procesami życiowymi, lecz nie abstrakcyjnym »życiem«" (Mayr, 2002, s. 20). Brak jasnej eksplikacji tego, co kryje się pod pojęciem życia, prowadzi do wielu niejasności i budzi pytania dotyczące tego, co decyduje o zakwalifikowaniu danego systemu do świata ożywionego. Dodatkowo mówienie o życiu wyłącznie w kategoriach autopojezy nie pozwala wyjaśnić oczywistych różnic poznawczych zachodzących pomiędzy organizmami oraz różnic w ich funkcjonowaniu, co w sposób satysfakcjonujący może zrobić psychologia porównawcza (por. Miłkowski, 2018a). Pytanie o definicję życia jest palące, ponieważ (1) można wskazać na istnienie form ewolucyjnych (np. wirusy) oraz nieewolucyjnych (sztuczne życie, roboty, sztuczna inteligencja itd.), których nie da się jednoznacznie zakwalifikować do układów żywych (por. Heller, Lubański, Ślaga, 1997, s. 299, 308; Latawiec, 1992); oraz (2) dyskutowana tutaj propozycja zbliża się do koncepcji funkcjonalistycznych, które w pewien sposób abstrahują od substratu materialnego życia (cielesności, morfologii itd.) oraz ujmują życie przede wszystkim w kategoriach przetwarzania informacji (por. Poczobut, 2011, s. 233) lub energii. O tym drugim powiem za chwilę. Zwolennicy teorii autopojetycznej oczywiście *explicite* wskazują na istotny wpływ ciała oraz usytuowania w środowisku, jednak proponowane przez nich ujęcie życia *implicite* nie korzysta z wartości eksplanacyjnej tych pojęć. Mianowicie: nie jest jasno określone, w jaki sposób koc Markowa charakteryzujący żywy organizm jest ucieleśniony. Na czym miałoby polegać faktycznie domknięcie operacyjne układu, który w pewnym sensie jest odgraniczony od środowiska przez granice własnego ciała? Jaką funkcję pełni morfologia ciała w aktywnym wnioskowaniu? Na te pytania omawiani autorzy nie dają jasnej odpowiedzi[25].

[25] Co prawda, wspominają o tym, że ciało współkonstytuuje model generatywny (Kirchhoff, Froese, 2017, s. 14), ale poza samą tą deklaracją nie wyjaśniają, na czym owa konstytucja miałaby polegać. A jest to istotny brak dla stanowiska, które chce uwypuklić znaczenie życia dla zrozumienia funkcjonowania systemów ergodycznych. Chcę przez to powiedzieć, że brak konkretnych analiz roli ciała w minimalizowaniu energii swobodnej, kluczowej dla utrzymania życia, powoduje, że stanowisko Kirchhoffa i Froese nieznacznie różni się od stanowisk, które ignorowałyby zupełnie rolę cielesności w tych aspektach.

Można tutaj postawić analogiczny zarzut do tego, który sformułował William Ramsey (2007) wobec stanowisk określanych jako reprezentacjonistyczne. Jego zdaniem zbyt często i łatwo w kognitywistyce przypisuje

Wielu badaczy jest również zgodnych co do tego, że nie da się podać jednej ostatecznej definicji życia (monoatrybutywnej) i w zamian ograniczają się do formułowania definicji poliatrybutywnych, które wskazują na wiele różnych właściwości układów żywych. Do tych własności można zaliczyć: (1) istnienie programów genetycznych; (2) występowanie w strukturze organizmów żywych specyficznych makrocząstek (hormony, enzymy itd.), które odpowiadają za ich rozwój i funkcjonowanie; (3) określoną organizację funkcjonalną; (4) określony rząd wielkości; (5) istnienie cykli życiowych i rozrodczych; (6) występowanie złożonych mechanizmów regulacyjnych, w tym sprzężeń zwrotnych; oraz (7) otwartości energetycznej i informacyjnej na oddziaływania ze środowiskiem. Zwolennicy ujęcia autopojetycznego zdają się literalnie ograniczać do tych dwóch ostatnich własności, co pozwala zaliczyć ich podejście do monoatrybutywnego traktowania życia, przy oczywistym ignorowaniu pozostałych własności, na które zwracają uwagę zwolennicy ujęć poliatrybutywnych.

Proponowane przez Kirchhoffa i Froese powiązanie teorii autopojetycznej z FEP zbliża ich koncepcję do podejść cybernetycznych. Dzieje się tak, ponieważ, o czym już wspomniałem, ujmują oni życie w kategoriach samoorganizacji, optymalizacji oraz wymiany energetyczno-informacyjnej ze środowiskiem. Powołują się przy tym na pogląd Fristona, który głosi, że system biologiczny „nie posiada modelu swojego świata – on jest tym modelem" (Friston, 2013c, s. 32). Friston, wypowiadając te słowa, korzysta ze sformułowania ukutego przez Rogera C. Conanta i W. Rossa Ashby'ego, że „każdy dobry regulator danego systemu musi być modelem tego systemu"[26] (Conant, Ashby, 1970; por. Friston, Thornton, Clark, 2012, s. 2; Friston i in., 2014, s. 5; Ramstead i in., 2018, s. 12). Samo podobieństwo formułowanych tutaj sądów nie wystarczy, aby ujęcie Kirchhoffa i Froese wiązać z cybernetyką, niemniej implikacje obu tych podejść są bardzo podobne. Zarówno cybernetyka, jak i teoria autopojezy traktuje życie w kategoriach samoodtwarzania i samoregulacji. Istotne jest również wytyczenie pomiędzy systemem a jego środowiskiem funkcjonalnej granicy, która wyznacza pewne minimum określające indywidualność i tożsamość tego systemu. Celem procesów życiowych, percepcyjnych i poznawczych, jest utrzymanie homeostazy i wewnętrznej organizacji, co w praktyce oznacza sprowadzenie ich do sterowania procesami energetycznymi

się reprezentacjom moce eksplanacyjne. Kluczowe jest to, czy przyjęcie reprezentacjonizmu pozwala lepiej wyjaśnić dane zjawisko lub ich klasę. Chodzi zatem o wykazanie, że używane w danej koncepcji czy modelu pojęcie reprezentacji nie jest trywialne, czyli że spełnia „wymóg opisu zadania" (*job description challege*), które polega na pokazaniu, w jakim sensie postulowane w ramach danej koncepcji czy modelu pojęcie reprezentacji rzeczywiście służy jako *reprezentacja* (Ramsey, 2007, s. XV). Sądzę, że wymóg sformułowany przez Ramseya można zastosować także do innych pojęć (np. ciała), którymi posługują się kognitywiści. W kontekście prowadzonych tutaj analiz należałoby zatem wykazać, czy pojęcie ciała pełni jakąś określoną eksplanacyjną funkcję, której nie dałoby się zredukować np. do funkcji realizowanych przez pojęcia organizmu albo systemu poznawczego. Twierdzę, że nie.

[26] Jest to tzw. twierdzenie o dobrym regulatorze, kluczowe dla cybernetyki. Por. także podrozdział *Normatywność homeostazy*.

(por. Seth, 2015, s. 7–8). Życie jest zatem podporządkowane pewnej teleonomii zachodzącej pomiędzy układem a jego odniesieniem (tj. środowiskiem). Mówienie o tym, że organizm jest modelem świata, jest niejasne. W cybernetyce ów fakt jest tłumaczony na podstawie izomorficznej relacji zachodzącej pomiędzy modelem a tym, co on modeluje. W ujęciu proponowanym przez zwolenników autopojezy takiego wyjaśnienia nie ma. Innego zresztą też.

Zależność pewnych rozwiązań ujęcia enaktywistyczno-autopojetycznego od cybernetyki staje się jeszcze wyraźniejsza, kiedy zwróci się uwagę na pojęcie umysłu, którym posługują się zwolennicy tego pierwszego. Jak wcześniej wspomniałem (podrozdział *Ujęcie enaktywiczne*), pojęcie to zostało przez Kirchhoffa i Froese zaczerpnięte z pracy Hutto i Myina i wzbogacone o pojęcie informacji w rozumieniu Shannona. Hutto i Myin twierdzą, że prosty umysł to ten, który jest pozbawiony treści, czyli nie zawiera żadnych reprezentacji. Jednak w przeciwieństwie do badaczy układów dynamicznych (por. Chemero, 2009, 2014) albo robotyki (por. Brooks, 1991) nie pokazują, w jaki sposób można modelować albo wyjaśnić funkcje umysłowe, nie korzystając jednocześnie ze słownika reprezentacjonistycznego. Pewne trudności wynikają także z utożsamienia poznawania z posiadaniem umysłu. Wystarczy przypisać pewnym artefaktom realizację procesów przetwarzania informacji (por. Dretske, 2004), aby uznać je za istoty mające proste umysły. Tutaj już dzieli nas krok od przypisania tym strukturom życia, co wiązałoby się z wykazaniem ciągłości pomiędzy posiadaniem prostych umysłów a byciem żywym organizmem. Wystarczyłoby zatem odwrócić rozumowanie Kirchhoffa i Froese. Zabieg ten jest dopuszczalny, ponieważ epistemiczny wymiar FEP, który podkreślają omawiani autorzy, wskazuje na konstytutywną zależność zachodzącą pomiędzy procesami życiowymi a procesami mentalnymi. Trudno się zatem zgodzić z Froese (2014), który twierdzi, że różnica pomiędzy żywym insektem a jego robotycznym modelem opiera się na posiadaniu prostego umysłu i autonomii[27]. Różnica pomiędzy tymi dwoma układami jest różnicą w przypisywaniu normatywności: normatywność, którą można przypisać żywemu owadowi, jest biologiczna; normatywność modelu jest zaś zależna od producenta czy architekta. Hutto i Myin nie wywodzą jednak z normatywności biologicznej normatywności weredycznej, czyli tej, która jest istotna z punktu widzenia wyjaśnienia procesów poznawczych (Hutto, Myin, 2013, s. 112)[28].

Podejście enaktywistyczne za sprawą rozumienia życia jako samoodtwarzania i samoregulacji, wsparte przez minimalne pojęcie prostych umysłów, pozwala je

[27] W inny, bardziej przekonujący, sposób argumentuje Margaret Boden, twierdząc, że różnica pomiędzy życiem biologicznym a życiem sztucznym zasadza się na realizacji procesów metabolicznych przez to pierwsze (Boden, 2000). Jednak i takie postawienie sprawy może budzić wątpliwości, bowiem można wykorzystywać mikrobiologiczne ogniwa paliwowe do projektowania sztucznych systemów metabolicznych (por. Ieropoulos, Melhuish, Greenman, 2003).

[28] Problematyczne jest także odniesienie się do (syntaktycznej) informacji Shannona jako odpornej na trudny problem treści (por. Miłkowski, 2015b).

traktować jako jedną z możliwych interpretacji podejścia cybernetycznego. Ranulph Glanville wprost stwierdza, że za sprawą wielu rozwiązań pojęciowych i definicyjnych należy traktować teorię autopojetyczną jako naturalne rozwinięcie tzw. nowej cybernetyki, która była odpowiedzią na krytykę pierwotnego stanowiska Conanta i Ashby'ego (por. Glanville, 2003; także Villalobos, 2013). Nowa cybernetyka twórczo włączyła w swoje badania takie zagadnienia jak autonomia czy samoodniesienie. W tym ujęciu „cykliczność jest traktowana poważnie" (Glanville, 2004). Porównanie podejścia enaktywistycznego z cybernetycznym nie jest samo w sobie zarzutem, niemniej pozwala ono zasadnie wątpić w moce eksplanacyjne tego pierwszego, postulowane przez jego zwolenników. Być może jest to jeden z powodów, dla którego nie zrewolucjonizowało ono nauk poznawczych ani nie dało mocnych podstaw eksperymentalnych, które pozwoliłyby w nowy sposób wyjaśnić wcześniej przebadane w innych podejściach zjawiska (por. Miłkowski, 2018a). Można zatem powiedzieć, że ujęcie enaktywistyczne oparte na teorii autopojezy ostatecznie proponuje (nie taki) nowy słownik dla już opisanych problemów, zaś w aspekcie spekulatywnym, który uważa się za charakterystyczny dla tego nurtu (np. problem ciągłości na linii życie – umysł), wciąż tkwi na poziomie pojęciowych dywagacji, konstruowania definicji projektujących[29] oraz prób coraz precyzyjniejszego wypowiedzenia swoich założeń bazowych. Tym można uzasadniać potrzebę uzupełnienia teorii Vareli i Maturany o FEP Fristona. W efekcie tego otrzymujemy redeskrypcje klasycznych tematów cybernetycznych, ostatecznie pozbawioną jakichkolwiek wyjaśnień (przyczynowo-nomologicznych, dynamicznych czy mechanistycznych). Trzeba się zgodzić z Miłkowskim, że podejście to należy

[29] Jako wyraźny przykład tej tendencji można podać pracę Chauncey'ego Mahera dotyczącą umysłów u roślin (Maher, 2017). Autor, aby uzasadnić hipotezę, która zakłada posiadanie prostych umysłów przez rośliny, konstruuje argument przeciwników tej hipotezy oraz odrzuca go, konstruując tym samym kontrargument wykorzystujący rozważania Hutto, Myina i zwolenników autopojezy. Argument prowadzący do odrzucenia życia mentalnego u roślin wygląda następująco (Maher, 2017, s. 106):

- posiadanie umysłu wymaga posiadania reprezentacji;
- rośliny nie reprezentują niczego; *zatem*
- rośliny nie mają umysłu.

Kontrargument Mahera (2017, s. 120):

- aby mieć umysł, wystarczy być systemem autopojetycznym i adaptacyjnym;
- rośliny są takimi systemami; *zatem*
- rośliny mają umysł.

Nietrudno zauważyć, że cały ten wywód opiera się na wielu definicjach projektujących („Posiadać umysł to posiadać reprezentacje"; „Posiadać umysł to być systemem autopojetycznym i adaptacyjnym"). Wystarczy zanegować to, co znajduje się w *definiensie*, aby zauważyć jawną fałszywość tych wyrażeń (przykładowo: „Posiadać umysł to nie posiadać reprezentacji" – zdanie to odniesione do wszystkich systemów poznawczych jest jawnie fałszywe, ponieważ ludzie oraz część zwierząt realizują swoje procesy poznawcze (przynajmniej ich część) na podstawie reprezentacji: pojęciowych, symbolicznych, wyobrażeniowych itd.). Można oczywiście bronić tego ujęcia, twierdząc, że wyznacza ono warunki minimalne, tzn. konieczne, ale niewystarczające do posiadania umysłu, niemniej wciąż problematyczny jest status tego rodzaju definicji oraz wyjaśnień budowanych na ich podstawie. Trudno byłoby bowiem bronić tezy, że są to definicje sprawozdawcze, chociażby z powodu ciągłych sporów wśród badaczy dotyczących tego, jak w ogóle rozumieć życie. Maher zaznacza, że w swoich analizach opiera się na sformułowanej przez Thompsona tezie „minds in life", która brzmi „wszystkie żyjące rzeczy (*things*) posiadają umysł" (Maher, 2007, s. 121). Teza ta jednak znowu jest zawoalowaną definicją projektującą o postaci: „Żywe jest to, co posiada umysł".

traktować jedynie jako tradycję badawczą. Wydaje się zatem, że słabości, za które krytykowano cybernetykę, są także obecne w koncepcjach inspirowanych teorią autopojetyczną.

Ostatnim zarzutem, który można postawić temu podejściu, jest pytanie o zawarty w nim *implicite* redukcjonizm metodologiczny. Redukcjonizm metodologiczny zakłada, że w miarę postępu badawczego prawa poszczególnych nauk da się zredukować do praw bardziej podstawowych. W tym przypadku prawa czy wyjaśnienia z zakresu nauk psychologicznych miałoby się sprowadzić do pojęć i wyjaśnień z zakresu teorii autopojetycznych, czy mówiąc bardziej oględnie – biologii. Po pierwsze nie jest jasne, dlaczego zjawiska badane przez psychologię lub neuronauki miałyby być wyjaśniane przez jedno autopojetyczne ujęcie; po drugie zaś samo to podejście, jak dotąd, nie może się pochwalić żadnymi eksperymentami oraz faktycznymi wyjaśnieniami. Wciąż jest to program w „budowie" i jak dotąd nie doprowadził do żadnej rewolucji, którą wieścili niektórzy z jego zwolenników (por. podrozdział *Enaktywizm*). Z tych powodów twierdzę, że entuzjazm Kirchhoffa, Froese i innych nie jest uzasadniony.

3.4.3. Ujęcie kognitywistyczne: argumenty

Przypomnijmy: podejście kognitywistyczne w odniesieniu do problemu ciągłości zachodzącej pomiędzy życiem a umysłem opiera się na przekonaniu, według którego takiej ciągłości nie ma. Umysłu nie da się wyjaśnić w kategoriach biologicznej samoorganizacji i autonomii, jak twierdzą niektórzy zwolennicy nonkognitywizmu i teorii autopojezy. Powodem tego ma być, przypomnijmy, przekonanie, że wyjaśnienie umysłu i procesów poznawczych wymaga odwołania się do jakiejś koncepcji informacji semantycznej. Jak wiemy, zdaniem Kirchhoffa i Froese, do wykazania przejścia pomiędzy życiem a prostą umysłowością wystarczy pojęcie informacji syntaktycznej. Mianowicie: koncepcja Fristona nie przesądza jednoznacznie, w jaki sposób należy rozumieć energię, która ma zostać zminimalizowana. Kirchhoff i Froese (2017, s. 6–7) utrzymują, że istnieją trzy możliwe i dopuszczalne interpretacje tego pojęcia:

1. Termodynamiczna: system zamknięty termodynamicznie to taki, jaki wchodzi w relacje ze swoim otoczeniem tylko dzięki pracy i wymianie cieplnej. W tym ujęciu informacja ma być różnicą pomiędzy dwoma rozróżnialnymi stanami (np. niskim i wysokim potencjałem grawitacyjnym). Minimalizacja tak rozumianej energii jest kluczowa dla systemu, bowiem ten, aby przetrwać, musi unikać (zgodnie z drugą zasadą termodynamiki) stanu równowagi termodynamicznej, czyli śmierci cieplarnianej.

2. Shannonowska: wiąże się z rozumieniem energii, które Friston określa mianem „teoretyczno-informacyjnego". Tutaj informacja jest miarą średniego rozkładu prawdopodobieństwa dotyczącą różnicy w komunikacie przesyłanym od nadawcy do odbiorcy (por. Shannon, 1948). Minimalizacja energii swobodnej odnosi się w tym ujęciu do minimalizacji ilości zaskoczenia i zakłada ona rozumienie informacji jako syntaktycznej (por. Friston, Thornton, Clark, 2012, s. 130).

3. Semantyczna: informacje należy rozumieć semantycznie. Oznacza to, że informacja nie tylko dotyczy ilości komunikatu, ale także ma swoją treść. Takie rozumienie informacji, zdaniem Kirchhoffa i Froese, należy wiązać z komputacjonizmem, reprezentacjonizmem i konserwatywnym ujęciem PP, wedle którego model generatywny jest reprezentacyjny (por. Gładziejewski, 2016, 2017b; Hohwy, 2013; Kiefer, Hohwy, 2018; Wiese, 2016a). Tutaj umysł jest nieustannie zaangażowany w obliczeniowe procesy przetwarzania informacji semantycznej dążące do przewidywania charakteru i treści informacji pochodzącej z wejść zmysłowych.

Rozumienie informacji (1) należy odrzucić, bowiem sam Friston nieustannie powtarza, że energia swobodna nie jest energią w sensie termodynamicznym. Spór zatem dotyczy tego, czy informacja w FEP jest syntaktyczna (2), czy semantyczna (3). Znamy już argumenty na rzecz interpretacji (2). Zastanowię się teraz, jakimi argumentami dysponują zwolennicy ujęcia semantycznego czy kognitywistycznego. Nie będę jednak wchodził w dyskusję dotyczącą tego, czy model generatywny należy interpretować w kategoriach reprezentacyjnych, czy nie. Jest to problem, który nie jest istotny dla przeprowadzanych tutaj analiz. Obecnie moim celem jest odpowiedź na pytanie o to, czy informacja syntaktyczna jest wystarczająca do wyjaśnienia funkcjonowania organizmów żywych? Następnie zastanowię się, czy FEP jest taką zasadą, jaka może wyjaśnić funkcje poznawcze systemów poznawczych i czy w ogóle może wyjaśnić cokolwiek. W tym celu odwołam się wpierw do propozycji Hohwy'ego, który jest zwolennikiem zastosowania FEP w PP, a następnie odniosę się do krytyki dotyczącej statusu eksplanacyjnego tej zasady. Przejdźmy zatem do argumentów za semantycznym charakterem FEP[30].

Argument z niewystarczalności wyjaśnień opartych na informacji syntaktycznej

Kirchhoff i Froese bronią swojego stanowiska, wykorzystując tezę Godfreya-Smitha i Sterelnego, że informację biologiczną należy wiązać z informacją Shannona (por. Godfrey-Smith, Sterelny, 2007). Charakterystyczne dla tych autorów

[30] Ograniczam się tutaj wyłącznie do dyskusji z zakresu filozofii biologii, które wiążą się z problemem wyjaśnienia zjawiska życia i odpowiedzią na pytanie o to, czy informacja syntaktyczna jest satysfakcjonującą eksplanacyjnie dla tego zagadnienia. Nie dyskutuję tutaj zatem sformułowanego przez Hutto i Myina (2013) trudnego problemu treści i jego rozwiązania.

jest „redukcyjne" podejście do analizy życia, które powiązałem z ujęciami monoatrybutywnymi. Ma to określone konsekwencje dla ich rozumienia informacji biologicznej. W artykule pt. *Information in biology* (2007) Godfrey-Smith stwierdza, że pojęcie informacji w biologii jest kluczowe dla zrozumienia bardzo wielu zjawisk i mechanizmów. Posługuje się nim genetyka, biologia rozwojowa czy teoria ewolucji. Najbardziej charakterystyczną rolą dla pojęcia informacji jest ta, która wiąże się z opisem relacji pomiędzy genami a różnymi procesami i strukturami. Zdaniem wielu biologów przyczynową rolę genów należy rozumieć w sensie „przenoszenia informacji". Informację biologiczną można tutaj interpretować na dwa sposoby: albo minimalistycznie, czyli jako informację syntaktyczną; albo w sposób mocniejszy i bardziej kontrowersyjny, czyli jako informację semantyczną. Z informacją shannonowską mamy do czynienia zawsze, gdy zachodzi jakaś korelacja czy zdarzenie warunkowe. Możemy zatem powiedzieć, że geny zawierają informację o wytwarzanych przez nie białkach, albo że zawierają one informację o fenotypie całego organizmu. Niemniej nie mówi się w ten sposób więcej niż wtedy, kiedy się mówi, że dym zawiera informację o ogniu albo pierścienie drzewa zawierają informację o jego wieku. Pojawia się jednak pewna wątpliwość dotycząca tego, czy używając takiego minimalnego pojęcia informacji, jesteśmy w stanie odróżnić biologię od innych nauk, przykładowo geologii?

Część autorów podkreśla, że informacja Shannona jest zbyt uboga, aby mogła wyjaśnić wiele zjawisk badanych w biologii. Należy do nich m.in. problem potencjalnych błędów. Mianowicie: informacja Shannona nie pozwala wyjaśnić, w jaki sposób dana zmienna może nieść fałszywą informację o innej zmiennej. Niektórzy twierdzą (por. Maynard-Smith, 2000; Sterelny, Smith, Dickison, 1996), że aby wyjaśnić możliwość błędnej reprezentacji, należy wprowadzić naturalistyczne pojęcie funkcji jako efektu procesów naturalnej selekcji (por. także Bielecka, 2018). Mówienie o funkcjach pozwala wyjaśnić nieprawidłowości w działaniu określonych mechanizmów czy struktur. Przykładowo: funkcją serca jest pompowanie krwi, a nie wydawanie efektów akustycznych. Dzieje się tak, ponieważ jest to funkcja, która doprowadziła do tego, że serce jest faworyzowane przez naturalną selekcję (Wright, 1973). Bez pojęcia funkcji nie dałoby się wyjaśnić tego, że serce może nie działać[31]. Do wyjaśnienia błędnej reprezentacji lub dysfunkcji potrzebny jest zatem minimalny słownik normatywny, którego wyjaśnienia opartego na informacji syntaktycznej nie daje się sformułować. Część badaczy uważa, że podobna strategia „teleofunkcjonalna" może pomóc zrozumieć semantyczne własności genów, a także tych struktur biologicznych o właściwościach semantycznych[32]. Można oczywiście utrzymywać, że posiadanie danej funkcji wcale nie implikuje posiadania treści.

[31] Problemem normatywności funkcji zajmuję się w rozdziale *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego*, szczególnie w podrozdziale *Pojęcie funkcji*.

[32] Sterelny, Griffiths (1999, s. 104): „Wyróżniającym testem informacji intencjonalnej bądź semantycznej jest fakt, że mówienie o błędzie bądź błędnej reprezentacji ma sens". Smith (2000, s. 193): „W biologii błędna

Funkcja nóg do chodzenia wcale nie reprezentuje chodu (Godfrey-Smith, 2007, s. 108). Z drugiej strony Sterelny, Smith i Dickison (1996) twierdzą, że istnieje związek pomiędzy funkcjami ewolucyjnymi i semantycznymi w odniesieniu do niektórych struktur biologicznych: geny dzięki swoim funkcjom reprezentują wyniki, które mają wytwarzać. Dodają również, że inne, także niezależne, czynniki pełniące podobne funkcje rozwojowe mają również właściwości semantyczne (por. także Sterelny, Griffiths, 1999[33]). Zatem nie tylko geny, ale także czynniki pozagenetyczne powinno się scharakteryzować w kategoriach informacji semantycznej.

Podsumowując powyższy wątek, należy stwierdzić, że trwa spór dotyczący tego, czy i w jaki sposób należy wykorzystać pojęcie informacji semantycznej w biologii. Badacze są zgodni co do tego, że pojęcie informacji Shannona nakłada pewne warunki minimalne dla zastosowania informacji w wyjaśnieniu zjawisk, niemniej to dopiero informacja semantyczna może pozwolić na znaczny postęp teoretyczny. Nie jest zatem jasne, dlaczego jako biolodzy czy filozofowie biologii mielibyśmy rezygnować z pełniejszego wyjaśnienia badanych przez nas zjawisk. Może być oczywiście również tak, że część zjawisk da się wyjaśnić przy użyciu wyłącznie informacji Shannona, zaś wyjaśnienie innych zjawisk zakłada wykorzystanie informacji semantycznej. Przedwczesne jest jednak uznanie samej informacji syntaktycznej za w pełni satysfakcjonującą w odniesieniu do wymogu wyjaśnienia życia i zjawisk z nim powiązanych[34].

Argument z błędów poznawczych

Artemy Kolchinsky oraz David Wolpert (2018) twierdzą, że klasyczne shannonowskie pojęcie informacji syntaktycznej nie pozwala dobrze wyjaśnić tego, w jaki sposób i dlaczego otoczenie jest przyczynowo konieczne dla utrzymania istnienia danego systemu (żywego albo sztucznego), rozumianego tutaj jako jego dążenie do stanu niskiej entropii. Chodzi zatem o wyjaśnienie sposobu, w jaki system, unikając równowagi termodynamicznej, jest jednocześnie podatny na różne kontrfaktyczne interwencje ze strony środowiska. Kolchinsky i Wolpert twierdzą, że dobre wyjaśnienie korelacji system – środowisko jest możliwe, o ile skorzysta się z pojęcia informacji semantycznej. Tę ostatnią rozumieją jako taką

reprezentacja jest możliwa, ponieważ istnieje zarówno rozwinięta struktura niosąca informacje, jak i rozwinięta struktura, która je otrzymuje".

[33] „Informacja intencjonalna wydaje się lepszym kandydatem na zrozumienie sposobu, w jaki geny przenoszą informacje rozwojowe i nic innego nie robią" (Sterelny, Griffiths, 1999, s. 104).

[34] Alistair Isaac w artykule *The semantics latent in Shannon information* (2019) powiada, że w informacji syntaktycznej są już *implicite* zawarte pewne intuicje dotyczące semantycznej wartości komunikatu. Shannon wprost twierdził, że zaproponowana przez niego teoria informacji nic nie mówi na temat znaczenia czy treści informacji, ponieważ nie jest to problem istotny z punktu widzenia inżyniera (Shannon, 1948, s. 31). Nie oznacza to jednak, że informacja w komunikacie nie ma treści.

informację, jaką system fizyczny musi mieć o swoim środowisku, aby utrzymać się przy istnieniu. Dodają przy tym, że należy jednak skonstruować taką teorię, jaka będzie odporna na zarzut głoszący, że informacja semantyczna ma znaczenie jedynie dla obserwatora czy badacza danego systemu, nie zaś dla tego systemu. Aby uniknąć tego zarzutu, należy skonstruować takie pojęcie informacji semantycznej, jakie jest wyłącznie oparte na wewnętrznej dynamice systemu, niezależnie od jego uwarunkowań ewolucyjnych oraz pochodzenia[35].

Punkt wyjścia analiz Kolchinsky'ego i Wolperta jest podobny do punktu wyjścia teorii autopojezy. Mianowicie: systemom autonomicznym (takim jak żywe organizmy, ale także niektóre roboty), które utrzymują swoje istnienie w obrębie danego środowiska, nie można przypisać funkcji samoorganizacji z zewnątrz. Jej opis jest niezależny zarówno od obserwatora, jak i historii ewolucyjnej. Zdaniem zwolenników autopojezy opis owej samoorganizacji i autonomii nie wymaga korzystania z pojęć semantycznych. Wystarczy odnieść się do dynamicznego sprzężenia ze środowiskiem oraz bezpośredniej (pozbawionej treści) intencjonalności. Inaczej sprawę stawiają Kolchinsky i Wolpert. Twierdzą oni, że system autonomiczny, aby utrzymać się przy życiu, musi nieustannie aktualizować swoją wiedzę na temat otoczenia, czyli uzyskiwać określone, a nie dowolne, informacje na podstawie ciągłej obserwacji. W tym sensie prosta bakteria realizuje ten sam mechanizm przetwarzania informacji jak makler giełdowy: bakteria chemotaktyczna wyczuwa stężenie chemiczne w swoim środowisku i następnie podąża w jego kierunku, tym samym lokalizując źródło pożywienia i energii. Takie autonomiczne zachowanie odróżnia tego rodzaju bakterię od „pasywnych struktur" w stylu komórek Bernarda, które powstają jedynie w wyniku oddziaływań zewnętrznych (w tym przypadku podgrzewania danej cieczy od dołu). Makler również musi nieustannie zbierać określone informacje, aby szybko reagować na dynamiczne zmiany na giełdzie.

To, co odróżnia zachowania autonomicznych systemów od pasywnych, to możliwość popełnienia błędu czy pomylenia się (por. Bickhard, 2014, 2016). Owa możliwość pojawia się zawsze wtedy, gdy system wykorzystuje informację, która może „utrudnić" mu utrzymanie własnego istnienia. Oznacza to, że relacja semantyczna pomiędzy systemem a środowiskiem jest asymetryczna, w przeciwieństwie do informacji syntaktycznej, która jest symetryczna jako informacja wzajemna (por. Kirchhoff, Froese, 2018). Przykładowo zakłada się, że bakteria chemotaktyczna ma informacje o swoim środowisku, ale nie oczekuje się, iż środowisko naturalne będzie mieć jakieś informacje semantyczne o bakterii.

Podsumowując ten wątek, należy stwierdzić, że pełne wyjaśnienie samoorganizacji i autonomii systemów żywych oraz sztucznych wymaga zastosowania informacji

[35] Omawiani autorzy twierdzą, że koncepcjom teleosemantycznym można zarzucić, iż z perspektywy systemu nie jest istotne, czy danej funkcji można przypisać treść; oraz to, że ma ona zastosowanie tylko do takich systemów, jakie przechodzą naturalną selekcję. Nie można zatem wyjaśnić funkcji systemów sztucznych lub nieożywionych, jeżeli nie odniesie się do ich wytwórcy czy projektanta (Kolchinsky, Wolpert, 2018, s. 2).

semantycznej, która pozwala wykazać asymetryczność relacji zachodzącej pomiędzy tym systemem a jego środowiskiem. Utrzymanie przez system niskiej entropii zakłada zatem możliwość popełnienia błędu lub jego uniknięcia, które kolejno zwiększają lub zmniejszają stan równowagi termodynamicznej systemu z jego otoczeniem.

Argument z komunikacji

Nicholas Shea, Peter Godfrey-Smith oraz Rosa Cao (2018) zwracają uwagę na fakt, że wielu badaczy twierdzi, iż informacja syntaktyczna jest wystarczająca do wyjaśnienia interakcji i komunikacji prostych systemów[36], a mimo tego domyślnie korzystają z koncepcji, która jest bogatsza niż ta zaproponowana przez Shannona. Tendencja ta jest szczególnie dostrzegalna w kontekstach, w których liczy się nieprawdziwość informacji. Wspomniani autorzy proponują taką teorię informacji, która w sposób bardziej adekwatny zda sprawę z wymiaru informacyjnego w tych prostych modelach. W tym celu wprowadzają pojęcie treści funkcjonalnej, które oddaje stopień, w jakim różne stany świata są zaangażowane w wymianę informacyjną pomiędzy nadawcami i odbiorcami. Dany komunikat ma treść funkcjonalną wtedy, gdy równowaga danego systemu jest utrzymywana przez określone procesy selekcji. Oznacza to, że dla określonego sygnału istnieje takie zachowanie odbiorcy tego sygnału, jakie nie pojawia się w przypadku reakcji na inny sygnał. Oczywiście odbiorca może reagować w taki sam sposób również na inne sygnały, ale wykluczona jest możliwość, aby reagował on w ten sam sposób na wszystkie możliwe sygnały w danym środowisku (Shea, Godfrey-Smith, Cao, 2018, s. 1015–1016). Shea i współautorzy powołują się tutaj na pogląd Briana Skyrmsa, który głosi, że zawartość informacyjna sygnału to wektor zmian dokonanych przez sygnał w rozkładzie prawdopodobieństwa po stanach świata (Skyrms, 2010). W tym ujęciu sygnały nie tylko zawierają treści na temat stanów świata (w odniesieniu do dyspozycji nadawcy), ale także niosą informacje o dokonanych czynnościach odbiorcy (w odniesieniu do jego dyspozycji). Pogląd Skyrmsa, na co zwraca uwagę Godfrey-Smith, prowadzi do ważnego pytania dotyczącego tego, co odbiorca sygnału faktycznie otrzymuje od jego nadawcy? Na to pytanie można podać dwie możliwe odpowiedzi: (1) jeżeli odbiorca dowiaduje się, jaki świat prawdopodobnie jest, to jednocześnie może się dowiedzieć, jak zmieniły się prawdopodobieństwa; albo (2) jeżeli odbiorca nie dowie się na podstawie sygnału, jaki świat prawdopodobnie jest, to w reakcji na brak dobrej informacji może podjąć pewne działania (Godfrey-Smith, 2011). Można zatem stwierdzić, że analiza informacji nie tylko

[36] Tendencję tę dobrze wyrażają Carl T. Bergstrom i Martin Rosvall, twierdząc, że „Kiedy teoretycy informacji mówią o kodowaniu, to nie myślą oni o własnościach semantycznych. Wszelkie własności semantyczne są wpychane w książkę kodów, czyli interfejs pomiędzy strukturą źródłową i strukturą kanału, który dla teoretyków informacji jest tak samo interesujący, jak książka telefoniczna dla socjologów" (2011, s. 171).

dotyczy prawdziwości treści, którą ona przekazuje, ale także jej nieprawdziwości czy niepełności. Zdaniem Skyrmsa błędna informacja pojawi się wtedy, kiedy sygnał zmniejsza prawdopodobieństwo zajścia danego stanu rzeczy, albo wtedy, gdy to prawdopodobieństwo zwiększa.

Shea, Godfrey-Smith i Cao uważają, że pojęcie treści funkcjonalnej jest użytecznym narzędziem w analizie komunikacji w systemach sygnalizujących. Swoją tezę popierają kilkoma przykładami. Omówię najprostszy z nich (Shea, Godfrey-Smith, Cao, 2018, s. 1018–1019). Wyobraźmy sobie, że istnieje system sygnalizacji, który opiera się na dwóch równie prawdopodobnych stanach rzeczy, dwóch prostych sygnałach i dwóch działaniach. Istnieją różne możliwe strategie nadawcze i odpowiadające im strategie odbiorcze. Są to kombinacje zachowań nadawcy i odbiorcy, w których obie strony uzyskują tyle samo w każdym procesie. Dzieje się tak, ponieważ sygnały są wykorzystywane do korelacji działań odbiorcy z określonym stanem świata. W tym prostym przypadku nadawca wysyła sygnał M1, odpowiadając na stan świata S1. Następnie odbiorca wytwarza sygnał A1, odpowiadając na informację o S1 zawartą w M1 itd. Pomiędzy nadawcą i odbiorcą sygnału jest utrzymana równowaga na podstawie określonego stanu świata. Treść funkcjonalną sygnału (M1) określa się poprzez zbadanie zachowania odbiorcy, które jest specyficzne dla tego, a nie innego sygnału (A1) w odniesieniu do określonego stanu rzeczy (S1), który jest warunkiem utrzymania stanu równowagi. Stabilizacja systemu jest oparta na wymianie informacyjnej pomiędzy nadawcą a odbiorcą w odniesieniu do określonych (probabilistycznych) stanów świata.

Powyższa analiza wzmacnia powszechne przekonanie głoszące, że mózg ludzki jest dobrym przykładem systemu sygnalizacyjnego, który ma zarówno własności semantyczne, jak i reprezentacyjne. Neurony lub całe ich populacje są nadawcami i odbiorcami informacji przenoszonych w potencjale czynnościowym. Cao (2012) uważa jednak, że dany system, w tym przypadku mózg, może przetwarzać faktyczną informację semantyczną, jeżeli spełni dwa wymagania: (1) odbiorca informacji musi być w stanie czerpać korzyści ze swojego otoczenia na podstawie sygnałów, które otrzymuje; oraz (2) musi mieć on pewną elastyczność wyboru reakcji w stosunku do odbieranego sygnału. Zdaniem Cao, jeżeli odniesie się te wymogi do sygnalizacji na poziomie pojedynczego neuronu lub nawet całej ich populacji, to okazuje się, że w najlepszym wypadku mamy tutaj do czynienia jedynie z bardzo słabą sygnalizacją w sensie semantycznym, o ile w ogóle się ona pojawi. Raczej pojawia się bogata ilościowo informacja w sensie korelacji. Taki wniosek z pewnością spełnia oczekiwania i wspiera argumentacje takich badaczy jak Kirchhoff i Froese czy Hutto i Myin. Niemniej, na co stanowczo zwraca uwagę Cao, prawdziwa reprezentacja świata i aktywność systemu w nim opiera się dopiero na odbiorze bogatej w treść informacji, która angażuje cały organizm, do którego konkretny zestaw neuronów wnosi bardzo niewielki wkład.

Omówione przeze mnie trzy argumenty pozwalają podać w wątpliwość przekonanie nonkognitywistów o tym, że sama informacja syntaktyczna w sposób adekwatny i niekontrowersyjny pozwala wyjaśnić ciągłość zachodzącą pomiędzy życiem a umysłem. Argumenty te wzmacniają wiele zarzutów, które sformułowałem podczas omawiania propozycji Kirchhoffa i Froese. Rzucają one również cień na zagadnienie, któremu poświęcony jest ten rozdział, czyli pytanie o moc eksplanacyjną fristonowskiego FEP i jego zastosowanie w analizach prowadzonych w ramach PP. Zanim podam ostateczną odpowiedź na to pytanie, omówię jeszcze propozycję Hohwy'ego, który z entuzjazmem odnosi się do tej ostatniej kwestii, a następnie porównam stanowisko kognitywistyczne z nonkognitywistycznym.

3.4.4. Ujęcie kognitywistyczne: propozycja Hohwy'ego

Zgodnie z FEP podstawową funkcją mózgu jest minimalizacja energii swobodnej (Friston, 2009, 2010). Organizm, aby żyć, *musi* minimalizować swoją energię swobodną, czyli sumę błędów predykcyjnych, które zwiększają stopień jego zaskoczenia informacją pochodzącą ze środowiska. Mózg zatem minimalizuje swoją energię swobodną analogicznie do tego, jak serce pompuje krew. Minimalizacja energii i pompowanie krwi to są funkcje kolejno mózgu i serca. To porównanie zdaniem Hohwy'ego jest jak najbardziej na miejscu, ponieważ identyfikacja funkcji danego organu pozwala na potraktowanie jej jako jednoczącej zasady organizującej pracę danego narządu. Serce inaczej pracuje podczas wypoczynku, a inaczej w trakcie wytężonej pracy. Niemniej za każdym razem pompuje krew. Z mózgiem jest podobnie: niezależnie od tego, co robi dany organizm, to mózg wciąż i bez spoczynku minimalizuje energię swobodną (Hohwy, 2015a, s. 2). Takie ujęcie jest redukcjonistyczne z dwóch powodów: (1) wszystkie dodatkowe funkcje mózgu, czynności, które on wykonuje, są redukowalne do jednej podstawowej funkcji; (2) wszelkie funkcje mózgu są sprowadzalne do zbioru prostych fizycznych mechanizmów. W tym drugim przypadku mamy do czynienia z redukcją w sensie ontologicznym (por. Searle, 1999, s. 159). Innymi słowy, umysł, jego funkcje i własności ostatecznie da się wyjaśnić funkcją minimalizowania energii swobodnej realizowaną przez organ fizyczny, jakim jest mózg.

Mózg musi minimalizować swoją energię swobodną, ponieważ optymalne dla danego organizmu jest funkcjonowanie w możliwie mało zaskakującym dla niego środowisku. Pojęcia zaskoczenia (*surprise*) nie należy bezpośrednio wiązać z jego potocznym odpowiednikiem, który pojawia się w takich zwrotach jak „wprawić kogoś w zdumienie" lub „zakłopotać kogoś czymś niespodziewanym". Przykładowo ryba wyjęta z wody narażona jest na większe zaskoczenie niż ryba pływająca w wodzie. Zaskoczenie jest zatem zawsze zrelatywizowane do danego modelu lub

zbioru oczekiwań[37]. Znajdowanie się poza wodą nie jest zupełnie zaskakujące dla człowieka albo kreta. Zaskakujące stany należy zatem rozumieć jako tę informację pochodzącą z wejść zmysłowych, która jest zaskakująca dla danego organizmu. Organizm zatem dąży do tego, aby otrzymać takie informacje, które będą zgodne z jego oczekiwaniami. Jak już wiemy, mózg nie dysponuje bezpośrednią informacją dotyczącą stanów środowiska, w którym żyje dany organizm (nie jest znany ich rozkład prawdopodobieństwa). Mózg zatem musi zbudować taki model generatywny świata, jaki będzie w sposób skuteczny obliczał rozkład prawdopodobieństwa jego stanów wewnętrznych i zewnętrznych na podstawie rozkładu prawdopodobieństwa informacji, która do niego dociera. Model zatem dotyczy oczekiwanych stanów zmysłowych, które są definiowane na podstawie jego wewnętrznych stanów intero- i eksterocepcyjnych (por. Hohwy, 2014). Model generatywny minimalizuje zatem swoje zaskoczenie na drodze minimalizacji różnicy pomiędzy prawdopodobieństwem wygenerowanej hipotezy a późniejszym rozkładem prawdopodobieństwa tej hipotezy w świetle danych zmysłowych i w kontekście modelu[38].

Minimalizacja energii swobodnej na drodze zmiany struktury sygnału zmysłowego (wnioskowanie aktywne) lub zmiany parametrów modelu (wnioskowanie percepcyjne) zwiększa jego dokładność w generowaniu predykcji. Zmiany te są możliwe tylko wtedy, powiada Hohwy, kiedy model może rekapitulować przyczynową strukturę świata, czyli wyjaśnić (na drodze wnioskowania do najlepszego wyjaśnienia) określone błędy predykcyjne. W ten sposób staje się on „zwierciadłem świata" (Hohwy, 2015a, s. 8). Minimalizacja energii swobodnej opiera się zatem na możliwości reprezentowania świata przez model w taki sposób, aby organizm mógł w nim działać. Zadanie reprezentowania stanów świata jest jednak, co podkreśla Hohwy, podporządkowane wymogowi homeostazy. Zatem mózg jest reprezentacyjny, bo minimalizuje swoją energię swobodną, a nie dlatego, że reprezentowanie świata jest warunkiem minimalizacji energii swobodnej[39]. Jest to istotna różnica, bowiem pozwala ona tłumaczyć reprezentacyjną naturę umysłu poprzez odniesienie

[37] Pomiędzy zaskoczeniem a energią swobodną zachodzi subtelna, ale istotna różnica. Aby ją pokazać, należy powołać się na istnienie koców Markowa. Przypomnijmy: koc Markowa ustanawia warunkową niezależność stanów wewnętrznych organizmów od stanów zewnętrznych środowiska. Te pierwsze są otwarte na drugie tylko w sensie warunkowym. Następnie w kocu Markowa można wyróżnić stany zmysłowe i aktywne. Odróżnia się je następująco: stany wewnętrzne nie mogą wpływać na stany zmysłowe, natomiast stany zewnętrzne nie mogą wpływać na stany aktywne (por. Friston, 2012). Taka warunkowa niezależność pozwala na statystyczne rozdzielenie stanów wewnętrznych koca (np. komórki albo organizmu) od stanów zewnętrznych jego środowiska. Stany wewnętrzne zatem na drodze minimalizacji zaskoczenia przez organizm kodują rozkład prawdopodobieństwa dotyczący stanów zewnętrznych, czyli przyczyn wrażeń zmysłowych odbieranych przez dany koc.

Energia swobodna jest zatem funkcją, która opisuje rozkład prawdopodobieństwa zakodowany przez wewnętrzne stany koca Markowa. Zaskoczenie jest funkcją stanów samego koca Markowa. Innymi słowy, energia swobodna jest funkcją przekonań probabilistycznych, zakodowanych przez stany wewnętrzne o stanach zewnętrznych (tj. oczekiwań dotyczących prawdopodobnych przyczyn bodźców sensorycznych) (Ramstead, Badcock, Friston, 2018, s. 3–4).

[38] Jest to tzw. dywergencja Kullbacka-Leiblera (dywergencja KL), która przyjmuje dwie wartości: albo wynosi ona zero, gdy nie ma rozbieżności; albo jest pozytywna, kiedy pojawia się rozbieżność, czyli błąd predykcyjny.

[39] Konserwatywne PP zostaje w tym ujęciu wzmocnione poprzez odniesienie do FEP.

do mechanizmów selekcji i historii ewolucyjnej. Energia jednak, o czym należy pamiętać, jest zawsze minimalizowana w odniesieniu do określonego modelu świata i jego własnych oczekiwanych stanów. Aby móc zatem wyjaśnić, dlaczego mózg działa w taki, a nie inny sposób, w jaki sposób może reprezentować świat, co jest warunkiem skutecznego działania organizmu w środowisku itd., należy odnieść się do eksplanacyjnego charakteru FEP, które wyjaśnia jego najbardziej podstawową funkcję, czyli minimalizację energii swobodnej.

Hohwy broni FEP przed potencjalnymi zarzutami, opierając się na analogii do teorii ewolucji (Hohwy, 2015a, s. 9–11). Powiada, że FEP przez swój bardzo ogólny charakter ma tendencję do monopolizowania wyjaśnień. Dla wielu może to być jej słabość, ale podobny zarzut można i nawet należałoby postawić teorii ewolucji, której nikt w świecie naukowym nie podważa. Teoria ewolucji podobnie jak FEP wskazuje na jeden podstawowy mechanizm, który ma dotyczyć wszelkich żywych organizmów. Może być tak, twierdzi Hohwy, że teoria ta wyjaśnia tylko 70% organizmów żywych, że należałoby wskazać jeszcze jakieś inne mechanizmy. Niemniej tego ograniczenia nikt nie powinien uznać za falsyfikujące wartość eksplanacyjną całej teorii. Wydaje się jednak, że raczej jest tak, iż teoria ewolucji albo wyjaśnia powstanie wszelkich organizmów żywych, albo nie wyjaśnia niczego. Nie ma drogi pośredniej. Podobnie ma się sprawa z FEP. Hohwy przewiduje, że FEP dopiero nabiera znaczenia w badaniach nad mózgiem i że z czasem stanie się tak powszechny jak powszechna jest metoda obrazowania mózgu, która z początku była traktowana z dużym sceptycyzmem (Hohwy, 2015a, s. 11).

Autor *The predictive mind* zwraca uwagę także na zarzut *trywialności*. Mianowicie: teoria jest *trywialna*, jeżeli *wszystko* za jej pomocą da się wyjaśnić; lub odwrotnie – *niczego* za jej pomocą nie da się wyjaśnić. Na ten zarzut Hohwy odpowiada, odnosząc się do kontekstów, które doprowadziły do sformułowania tej zasady, i które stanowią ramę uzasadniającą FEP. Są to m.in. filozoficzne analizy mózgu jako maszyny predykcyjnej (por. Ibn al-Hajtam; Helmholtz, 1867); wcześniejsze (obliczeniowe, teoretyczne i informacyjne) modele funkcjonowania mózgu kładące nacisk na istnienie połączeń wstecznych i sprzężeń zwrotnych, które były falsyfikowane formalnie i empirycznie; odniesienie do znaczenia mechanizmów przewidywania dotyczących reprezentacji świata i poziomu precyzji sygnału itd. Zasada energii swobodnej (FEP) okazałaby się z gruntu fałszywa, gdyby udało się wykazać, że istnieje takie żywe stworzenie, jakie nie podejmowałoby działań mających na celu utrzymanie stanu homeostazy. Takie stworzenie nie zmieniałoby swoich stanów wewnętrznych względem środowiska i zapewne bardzo trudno byłoby o nim mówić jako o żywym.

Hohwy nie tylko wierzy w powszechny charakter wyjaśnień oferowanych przez FEP, ale dodatkowo twierdzi, że jest to zasada kompatybilna z mechanistycznymi modelami wyjaśnień w naukach o poznaniu. Nie wchodząc w szczegóły, należy powiedzieć, zdaniem Hohwy'ego, że funkcjonalność FEP jest kompatybilna z wy-

maganiami stawianymi przez mechanicystów i może doprowadzić badaczy do wielopoziomowych wyjaśnień mechanistycznych.

3.4.5. Ujęcie nonkognitywistyczne a ujęcie kognitywistyczne

Przekonanie o dużej mocy eksplanacyjnej FEP łączy ujęcia nonkognitywistyczne i kognitywistyczne. Zwolennicy zarówno pierwszego, jak i drugiego podejścia podkreślają unifikujący, prymarny i konieczny wymiar FEP jako zasady wyjaśniającej funkcjonowanie organizmów żywych. Ich badania dotyczą głównie wyjaśnień na prostych poziomach życia oraz poznania, niemniej zarówno jedni, jak i drudzy są przekonani co do tego, że FEP może wyjaśnić wiele innych zjawisk. Fakt ten sam Friston i jego współpracownicy wielokrotnie podkreślają: FEP jest zasadą, która ma mieć zastosowanie do każdego biologicznego systemu – „od organizmów jednokomórkowych do sieci społecznych" (Friston, 2009, s. 293)[40]. Z tej perspektywy Kirchhoff i Froese, z jednej strony, oraz Hohwy, z drugiej, są głosicielami tezy, zgodnie z którą FEP nie tylko jest zasadą eksplanacyjną, ale także zasadą konstytutywną, bowiem nie tylko wyjaśnia organizację i funkcjonowanie struktur biologicznych, ale także wskazuje na mechanizm, który je konstytuuje, czyli mechanizm minimalizacji energii swobodnej.

To, co dzieli zwolenników nonkognitywizmu oraz kognitywizmu, to przede wszystkim inne rozumienie informacji oraz inne podejście eksplanacyjne. Obie te kwestie ściśle się ze sobą wiążą. Pierwszą różnicę omówiłem szczegółowo wcześniej. Tutaj ograniczę się do jej podsumowania. Zdaniem nonkognitywistów minimalizacja energii swobodnej w sensie fristonowskim odnosi się jedynie do informacji syntaktycznej. Informacja Shannona jest konieczna i wystarczająca do wyjaśnienia organizacji i funkcjonowania organizmów żywych oraz podstawowych form umysłowości. Założenie o eksplanacyjnej wystarczalności informacji syntaktycznej zakłada odrzucenie przekonania, zgodnie z którym informacja semantyczna (i powiązane z nią treść semantyczna oraz reprezentacje) pełni jakąś istotną funkcję eksplanacyjną. Można to rozumieć na dwa sposoby: (1) albo wyjaśnienia oparte na pojęciu treści lub reprezentacji nie wnoszą nic nowego do rozumienia życia i poznania; (2) albo pojęcia takie jak treść oraz reprezentacja nie mają żadnych realnych desygnatów i pełnić mogą w najlepszym wypadku funkcje instrumentalne (por. Egan, 2014), co w efekcie może także prowadzić do przyjęcia podejścia (1). Podejścia te można, za Chemero, wiązać kolejno z tzw. hipotezą dotyczącą wiedzy oraz hipotezą dotyczącą natury (Chemero, 2014)[41]. Dla potrzeb

[40] Zastanawiające jest potraktowanie sieci społecznych przez Fristona jako swoistego systemu biologicznego.

[41] Pierwsza z nich głosi, że „najlepsze modele i/lub wyjaśnienia naturalnych systemów poznawczych nie będą odwoływać się do jawnych reprezentacji. Tzn. nasze najlepsze modele i/lub wyjaśnienia systemów poznawczych nie będą zawierać nic, co stanowi reprezentacje zgodnie z naszą teleologiczną teorią reprezentacji";

niniejszych analiz określę je jako „antyreprezentacjonistyczną hipotezę dotyczącą wiedzy" oraz „antyreprezentacjonistyczną hipotezę dotyczącą natury". Hipoteza dotycząca natury ma charakter przedmiotowy, czyli orzeka coś na temat natury życia i (lub) umysłu, zaś hipoteza dotycząca wiedzy jest metaprzedmiotowa, bowiem dotyczy sposobu, w jaki powinno się uprawiać naukę, czy w interesującym nas kontekście – kognitywistykę. Trudno powiedzieć, pod którą z hipotez podpisują się zwolennicy nonkognitywizmu. Ich deklaracje w tym aspekcie nie są jednoznaczne, szczególnie, że nie głoszą oni globalnego antyreprezentacjonizmu: wszak wyższe funkcje poznawcze mogą w ich ujęciu mieć charakter reprezentacyjny. Można jednak śmiało powiedzieć, że kognitywizm w ujęciu Hohwy'ego odrzuca obie te hipotezy. Należy bowiem stwierdzić, że zdaniem Hohwy'ego wyjaśnienia reprezentacyjne pełnią istotną funkcję eksplanacyjną w kognitywistyce i neuronaukach, ponieważ systemy poznawcze są reprezentacyjne *per se*. Kognitywizm zatem nie tyle odrzuca obie te hipotezy, ile przeformułowuje je na modłę reprezentacyjną:

Reprezentacjonistyczna hipoteza dotycząca natury: naturalne systemy poznawcze korzystają z reprezentacji. Tzn. wewnątrz systemu poznawczego istnieją takie struktury, jakie spełniają warunki określane przez teleologiczną teorię reprezentacji.

Reprezentacjonistyczna hipoteza dotycząca wiedzy: najlepsze modele i (lub) wyjaśnienia naturalnych systemów poznawczych odwołują się do jawnych reprezentacji. To znaczy, że nasze najlepsze modele i (lub) wyjaśnienia systemów poznawczych zawierają to, co stanowi reprezentacje zgodnie z naszą teleologiczną teorią reprezentacji.

Jak nietrudno zauważyć podział na kognitywizm/nonkognitywizm pokrywa się w pewnym stopniu, z omówionym przeze mnie w rozdziale *Pojęcie przetwarzania predykcyjnego* podziałem na konserwatywne i radykalne PP. Już teraz można przyjąć twierdzenie, że zasadnicza różnica pomiędzy tymi stanowiskami dotyczy sposobu przyjęcia bądź odrzucenia powyższych hipotez. Zależnie od przyjętego podejścia można wyróżnić cztery grupy stanowisk:

(1) Zakłada się obie hipotezy reprezentacjonistyczne. Tym samym odrzuca się obie hipotezy antyreprezentacjonistyczne. Stanowisko to określę jako „mocny reprezentacjonizm".

(2) Zakłada się reprezentacjonistyczną hipotezę wiedzy przy jednoczesnym odrzuceniu hipotezy dotyczącej natury. Tym samym odrzuca się tylko antyreprezentacjonistyczną hipotezę wiedzy przy jednoczesnym zachowaniu antyreprezenta-

druga zaś, że „naturalne systemy poznawcze nie korzystają z reprezentacji. Tzn. wewnątrz systemu poznawczego nie istnieje nic, co spełnia warunki określane przez teleologiczną teorię reprezentacji" (Chemero, 2014, s. 84). Rozszerzam w tym miejscu hipotezy sformułowane przez Chemero o wymiar biologicznego życia.

cjonistycznej hipotezy natury. Stanowisko to określę jako „słaby reprezentacjonizm" (instrumentalizm).

(3) Zakłada się reprezentacjonistyczną hipotezę natury przy jednoczesnym odrzuceniu hipotezy wiedzy. Tym samym odrzuca się tylko reprezentacjonistyczną hipotezę natury przy jednoczesnym odrzuceniu reprezentacjonistycznej hipotezy wiedzy. Stanowisko to określę jako „słaby antyreprezentacjonizm".

(4) Odrzuca się obie hipotezy reprezentacjonistyczne. Tym samym zakłada się obie hipotezy antyreprezentacjonistyczne. Stanowisko to określę jako „mocny antyreprezentacjonizm".

Jedynie stanowisko określone przeze mnie jako „mocny reprezentacjonizm" spełnia warunki, które pozwalają zaliczyć je do konserwatywnego ujęcia PP. Pozostałe stanowiska można w mniejszym lub większym stopniu wiązać z radykalnym PP. Do radykalnego PP należy również zaliczyć podejścia, które zakładają, że tylko niektóre poziomy modelu generatywnego mają charakter reprezentacyjny lub dają się wyjaśnić przy użyciu reprezentacji (por. Clark, 2015a). Takie stanowiska określę mianem „umiarkowanego reprezentacjonizmu" lub „reprezentacjonizmu hybrydowego":

(5) Zakłada się obie hipotezy reprezentacjonistyczne (lub tylko jedną z nich) w odniesieniu do *n*-liczby poziomów modelu generatywnego. Tym samym uznaje się obie hipotezy antyreprezentacjonistyczne (lub tylko jedną z nich) w odniesieniu do pozostałych poziomów modelu generatywnego.

Zarysowana przeze mnie mapa stanowisk w sporze o reprezentacje pozwala na wstępne określenie statusu ontologicznego i metodologicznego podejść kognitywistycznego i nonkognitywistycznego w odniesieniu do FEP. Kognitywizm spod znaku Hohwy'ego spełnia warunki mocnego reprezentacjonizmu. Można go zatem zaliczyć do konserwatywnego PP. Nonkognitywizm z pewnością zalicza się do radykalnej odmiany PP. Nie da się jednak, z powodu braku wyraźnych deklaracji jego zwolenników oraz niejasności metodologicznych, zaliczyć go jednoznacznie do któregokolwiek ze stanowisk w sporze o reprezentacje.

W tym miejscu dochodzimy do drugiej różnicy dzielącej zwolenników kognitywizmu i nonkognitywizmu, czyli różnicy w podejściu do natury wyjaśnień. Kirchhoff i Froese podkreślają (2017), że to, co odróżnia ich stanowisko od podejścia Hohwy'ego, to przekonanie, iż pomiędzy życiem a umysłowością istnieje bezpośrednie przejście. Sama ta teza nie jest do końca jasna, bowiem można ją rozumieć: albo jako (1) twierdzenie, że życie jest konstytutywne dla pojawienia się umysłu; albo jako (2) twierdzenie, że zjawiska życia i umysłowości można wyjaśnić, korzystając z tego samego modelu wyjaśnień. Niemniej wydaje się, że omawiany spór ostatecznie nie tyle dotyczy tych kwestii, ile różnic wynikających z przyjęcia innych założeń

dotyczących hipotez o wiedzy i naturze. Stanowisku nonkognitywizmu, powtórzę, można zapewne przypisać podejście, które określiłem jako „mocny antyreprezentacjonizm", niemniej w samym tekście Kirchhoffa i Froese taka deklaracja się nie pojawia[42]. Innym i istotniejszym problemem jest to, że nie wiadomo, jaką teorię wyjaśnień przyjmują zwolennicy nonkognitywizmu. W przypadku kognitywizmu Hohwy'ego sprawa jest jasna, bowiem sam autor *The predictive mind* powołuje się na neomechanicyzm i mechanistyczny model wyjaśnień naukowych (por. Hohwy, 2015a, 2015b). Z nonkognitywistami sprawa jest o tyle trudna, że powołują się oni na różnych autorów z zakresu filozofii kognitywistyki, zaś ci autorzy korzystają z różnych modeli eksplanacyjnych. Przykładowo Chemero jest zwolennikiem dynamicyzmu; Hutto i Myin przyjmują metodologię opartą na elementach behawioryzmu Skinnera czy instrumentalizmu Dennetta, zaś sam Friston pozostaje bardzo ambiwalentny, jeżeli chodzi o problem wyjaśnień: czasem powołuje się na mechanicyzm, a czasem na inne modele. Nie jest jasne również, o jakiego rodzaju modelowaniu jest mowa w FEP. Pewną wskazówką są inspiracje płynące ze strony teorii systemów autopojetycznych. Ta ostatnia bowiem kojarzona jest w filozofii nauki z organicyzmem. Na potrzeby niniejszych analiz powiążę zatem nonkognitywizm z organicyzmem, zaś kognitywizm z mechanicyzmem. Obecnie mechanistyczny model wyjaśnień przeciwstawia się głównie teorii układów dynamicznych, niemniej wśród zwolenników PP nie ma orędowników dynamicyzmu, czyli badaczy, którzy faktycznie modelowaliby czynności poznawcze przy użyciu teorii układów sprzężonych.

3.5. Trudności eksplanacyjne zasady energii swobodnej

Niniejsze rozważania będą poświęcone pytaniu dotyczącemu eksplanacyjnej wartości FEP i pewnym trudnościom teoretycznym z tą zasadą związanym. Zgodnie z FEP każdy system, który minimalizuje swoją energię swobodną (w sensie informacyjno-teoretycznym), opiera się naturalnej tendencji do fizycznej dezintegracji (Friston, 2010, s. 1; Friston, 2012, s. 2100; por. także Ashby, 1947; Kauffman, 1993). System zatem, aby żyć, musi minimalizować swoją energię swobodną, co w praktyce oznacza, że minimalizacja energii swobodnej jest warunkiem życia. FEP jest zasadą, która wyjaśnia organizację systemów biologicznych [od jednokomórkowców po sieci społeczne (Friston, 2009, s. 293)], które z definicji opierają się rozkładowi w czasie (Ramstead, Badcock, Friston, 2018, s. 2). Umożliwia ona wyjaśnienie nie tylko struktury i funkcji mózgu [m.in. organizacji korowej oraz potencjału aktywności (por. Friston, Kilner, Harrison, 2006, s. 70)] na podstawie

[42] W innych pracach Kirchhoffa stanowisko mocnego antyreprezentacjonizmu jest przyjmowane *explicite* (por. Kirchhoff, Robertson, 2018).

samego faktu istnienia organizmów (Friston i in., 2010, s. 1), ale także działania oraz zachowania danego organizmu (Friston, 2012, s. 2102). Wiążąc teorię informacji z tezą o adaptacyjnym charakterze życia biologicznego w terminach sieci bayesowskich i teorii optymalnej decyzji, umożliwia wyjaśnienie natury procesów decyzyjnych, kontroli motorycznej i kosztów informacyjnych (Friston, 2012, s. 2102), a także wielu aspektów anatomii i fizjologii określonych organizmów (Friston, 2009, s. 295). Swoim zasięgiem eksplanacyjnym FEP ma zatem obejmować m.in. następujące zagadnienia[43]:

- kategoryzacja percepcyjna (Friston, Kiebel, 2009);
- wnioskowanie percepcyjne (Kiebel, Daunizeau, Friston, 2009);
- mowa (Kiebel, Daunizeau, Friston, 2009);
- uczenie percepcyjne (Friston, Kiebel, 2009);
- mechanizmy uwagi (Feldman, Friston, 2010);
- obserwacje działań i neurony lustrzane (Friston, Mattout, Kilner, 2011);
- działania ukierunkowane na cel (Friston i in., 2009);
- systemy dynamiczne (Friston, 2010);
- morfologia komórkowa (Friston i in., 2015a);
- wartości epistemiczne (Friston i in., 2015b);
- przetwarzanie predykcyjne (Friston, Kiebel, 2009);
- życie (Friston, 2013c).

Nietrudno spostrzec, że FEP ma olbrzymie pretensje eksplanacyjne. Omówiony przeze mnie spór, dotyczący roli FEP w wyjaśnieniu ciągłości na linii życie – umysł, jest tylko drobnym przykładem zastosowania tej zasady. Zwolennicy FEP wielokrotnie podkreślają, że jest to teleologiczna reguła, która pozwala połączyć biologię z teorią informacji i tym samym wyjaśnić konstytutywną relację pomiędzy mózgiem a środowiskiem (Allen, Friston, 2018, s. 2476). FEP ma mieć status zasady fundamentalnej czy wręcz „obowiązkowej" w neuronauce (Friston, Kilner, Harrison, 2006, s. 71; Friston, Stephan, 2007). Zasada ta ma charakter normatywny, bowiem wyjaśnia, dlaczego mózg musi z konieczności być zaangażowany w ucieleśnione przetwarzanie predykcyjne (Allen, Friston, 2018, s. 2472). Jest to zasada normatywna także dlatego, że wykazuje, iż organizmy niejako muszą opierać się drugiej zasadzie termodynamiki, aby mogły utrzymać się przy istnieniu (Allen, Friston, 2018, s. 2473)[44].

Podsumowując: FEP w myśl jej zwolenników to pierwsza i normatywna zasada wyjaśniająca dążenie każdego żywego organizmu do utrzymania stanu homeostazy. Jest to zasada tautologiczna, bowiem orzeka ona tylko tyle, że organizm żyje, jeżeli może żyć. Tautologiczność FEP jest taką samą tautologicznością jak ta, która pojawia się w teorii selekcji naturalnej [„Dlaczego ja tu jestem? – Ponieważ mam

[43] Lista ta z całą pewnością nie jest kompletna.

[44] Zagadnieniem normatywności FEP zajmuję się w rozdziale *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego*.

sprawność adaptacyjną: Dlaczego mam sprawność adaptacyjną? – Bo tutaj jestem" (Friston, Thornton, Clark, 2012, s. 2)]. Truizmem jest stwierdzenie, że organizm żywy żyje, ponieważ może żyć, zauważa Friston. Nie jest to jednak truizm, który można zbagatelizować. Wymaga on wyjaśnienia. Zadaniem FEP jest zatem wyjaśnienie tego, w jaki sposób ten „truizm" jest realizowany (Friston, Thornton, Clark, 2012, s. 6; por. także Hohwy, 2015a, s. 11). Jako „zasada życia" FEP obejmuje swoim zasięgiem funkcjonowanie organizmów żywych od poziomu molekularnego aż po wyższe procesy poznawcze. Schematycznie realizację FEP na poszczególnych poziomach życia organizmów można przedstawić w następujący sposób:

- poziom fizyczny: unikanie entropii;
- poziom biologiczny: przetrwanie;
- poziom fizjologiczny: utrzymanie homeostazy;
- poziom informacyjno-teoretyczny: minimalizacja energii swobodnej oraz redukcja zaskoczenia;
- poziom poznawczy: minimalizacja błędów predykcyjnych, maksymalizacja dopasowania modelu do niepewnego środowiska.

W kontekście analiz przeprowadzonych w tej pracy konieczne jest teraz omówienie zastosowania FEP w badaniach nad PP, które już w tym rozdziale kilkakrotnie było anonsowane. Przypomnijmy: Friston powiada, że dla zrozumienia dynamiki i funkcjonalności mózgu kluczowe jest odniesienie do jego hierarchicznej organizacji (por. Felleman, Van Essen, 1991). Organizację tę można dobrze opisać w kategoriach hierarchicznych modeli inwersyjnych, które minimalizują błędy predykcyjne. Kategorie te opisują sposób, w jaki stany neuronalne samoorganizują się pod wpływem działania czynników zewnętrznych (Friston, Kiebel, 2009, s. 1213). W zgodzie z FEP stany wewnętrzne danego organizmu (w tym przypadku stany neuronalne mózgu) można definiować w kategoriach koców Markowa, które, aby zachować swoją integralność i autonomię, muszą minimalizować energię swobodną (w sensie informacyjno-teoretycznym), czyli różnicę pomiędzy oczekiwaniami dotyczącymi tego, jaki świat jest (czyli sposobem jego modelowania), a tym, jaki świat jest. Minimalizacja energii swobodnej jest kluczowa dla organizmu, bowiem mechanizm ten pozwala na redukcję stopnia niepewności i zaskoczenia. Mechanika bayesowska PP redukuje zaskoczenie poprzez śledzenie i minimalizowanie rozbieżności pomiędzy danymi przesyłanymi przez zmysły a kodowanymi odgórnie predykcjami [wiążę się to z asymetrią funkcjonalną przewidywań i błędów, która opiera się na współdziałaniu dwóch rodzajów komórek piramidalnych (por. Clark, 2013b; Park, Friston, 2013)], modyfikując ich precyzję na podstawie mechanizmów neuromodulacyjnych. Zwolennicy zastosowania FEP w PP sugerują, że minimalizację energii swobodnej należy bezpośrednio wiązać z bayesowską minimalizacją błędów predykcyjnych. Innymi słowy, energia swobodna zostanie zminimalizowana wtedy, gdy albo zostaną zmienione parametry modelu, czyli przekonania (*priors*)

oraz predykcje dotyczące źródeł i charakteru informacji zmysłowej (wnioskowanie percepcyjne), albo gdy nastąpi zmiana w środowisku, tzn. zmieni się struktura przyczynowa określonych stanów świata (wnioskowanie aktywne) (por. Badcock, Friston, Ramstead, 2019, s. 8; Friston, Daunizeau, Kiebel, 2009)[45].

Zdaniem Fristona i jego zwolenników każdy organizm można opisać w kategoriach (wielowymiarowych) hierarchicznie zagnieżdżonych koców Markowa. Koce te są zorganizowane czasoprzestrzennie, a swoją integralność zachowują, z jednej strony, dzięki minimalizowaniu energii swobodnej, a z drugiej, dzięki relacji pomiędzy kolejnymi poziomami hierarchii koców Markowa. Umożliwia to powstanie kolejnych poziomów (informacyjnych, fizycznych, chemicznych, biologicznych, psychologicznych i społeczno-kulturowych) organizacji. Hierarchicznie zagnieżdżone koce Markowa są warunkowane dynamiką doboru naturalnego i mechanizmów selekcji naturalnej (por. Badcock, Friston, Ramstead, 2019, s. 11) oraz prawami statystycznej termodynamiki. Dzięki temu zachowują swoją integralność, statystyczną niezależność od środowiska, samoorganizują się i długoterminowo minimalizują swoje zaskoczenie (Ramstead, Badcock, Friston, 2018, s. 5). W odniesieniu do PP należy stwierdzić, że kolejne poziomy modelu są implementowane w mózgu jako zorganizowane hierarchicznie koce Markowa, nieustannie zaangażowane w proces ważonej precyzją minimalizacji błędów predykcyjnych, na drodze minimalizacji energii swobodnej (Friston, 2009). Mówienie o zagnieżdżonych czasowo kocach Markowa ma lepiej wyjaśnić działania organizmu, które są reakcją na dynamiczne zmiany generowane w środowisku. Przykładowo ruch ręki składa się z mniejszych i bardziej elementarnych ruchów, które są strukturyzowane przez następstwo czasowe (Badcock, Friston, Ramstead, 2019, s. 10–11). Podobnie sprawa wygląda w odniesieniu do hierarchii modelu generatywnego, która jest konstytuowana czasoprzestrzennie.

Podsumowując ten wątek, należy stwierdzić, że zdaniem części badaczy FEP umożliwia wyjaśnienie mechanizmów PP poprzez zastosowania takich kategorii teoretycznych jak: energia swobodna, zaskoczenie, koce Markowa, wnioskowanie aktywne/percepcyjne; za pomocą modelowania Bayesa i metod stosowanych w neuronaukach (takich jak fMRI czy EEG) oraz również, jak twierdzą niektórzy, ram eksplanacyjnych wyznaczanych przez neomechanicyzm (por. Hohwy, 2015a, 2015b; także Friston, Kilner, Harrison, 2006). Co ciekawe, w najnowszych pracach napisanych wspólnie przez Karla Fristona, Maxwella Ramsteada i Paula Badcocka (por. 2018, 2019) wielokrotnie mówi się o mechanizmach i mechanistycznej organizacji umysłu. W pracy Badcocka i in. (2019) pojawiają się nawet odwołania do prac Carla F. Cravera i Williama Bechtela – teoretyków wyjaśnień mechanistycz-

[45] Pojęciem wnioskowania aktywnego szczegółowo zajmuję się w rozdziale *Mechanizmy normatywne a działania w przetwarzaniu predykcyjnym*.

nych[46]. Nie jest jednak zupełnie jasne, w jaki sposób FEP miałaby być wyjaśniana mechanistycznie (por. Colombo, Wright, 2018).

Modelowanie w nauce, najogólniej mówiąc, jest czynnością relacyjną. Przeważnie obejmuje ono określony wzorzec, szablon, konstrukcję czy matrycę (określane zwykle mianem „modelu") oraz wyrażane *explicite* i *implicite* pewne ramy teoretyczne i interpretacyjne (oparte na danych założeniach, definicjach, uogólnieniach itd.). Jest to kluczowa czynność wiedzotwórcza szczególnie w biologii i psychologii, a także w wielu dyscyplinach wchodzących w obręb kognitywistyki.

Można wyróżnić następujące rodzaje modeli (Barandiaran, Chemero, 2009):

- mechanistyczne: zakładają jedno-jednoznaczną relację przyczynową pomiędzy komponentami modelu a elementami wyjaśnianego zjawiska (rozumianymi w kategoriach mechanizmu) (por. Bechtel, 2008; Craver, 2007; Glennan, 2017; Kaplan, 2011)[47];
- funkcjonalne lub dynamiczne: odwołują się raczej do określonych zachowań czy działań (na zasadzie analizy funkcji wejście – wyjście) niżeli do istnienia określonej korespondencji pomiędzy modelem a modelowanym zjawiskiem; opisują czarno-skrzynkowo badane zjawisko (por. Chemero, 2014; Van Gelder, 1995);
- generyczne lub ogólne: korespondują z wieloma zjawiskami w sposób abstrakcyjny i ogólny (por. Griffith, Nersessian, Goel, 1996);
- pojęciowe: tworzone są na podstawie określonej teorii, symulują konkretne teoretyczne założenia i pojęcia, ilustrują ogólne zasady rządzące daną klasą zjawisk (por. Barandiaran, 2007).

Sam Friston i jego zwolennicy wprost nie deklarują, o jaki rodzaj modelowania w FEP chodzi, niemniej nieustannie podkreślają jej eksplanacyjny charakter. Docieramy do dość nietypowej sytuacji: FEP stwierdza, że organizmy dążą do stanu homeostazy, jednak nic nie mówi na temat faktycznych mechanizmów realizujących tę zasadę. Sprawę dodatkowo komplikuje sam Friston, kiedy stwierdza, że w odniesieniu do FEP należy odróżnić jej rozumienie jako zasady normatywnej od rozumienia jej jako teorii lub hipotezy dotyczącej tego, w jaki sposób ta zasada jest realizowana. Z tego powodu, jak zauważają Friston, Fortier, Friedman (2018, s. 21):

> FEP należy odróżnić od takich rzeczy jak kodowanie predykcyjne i hipoteza mózgu bayesowskiego. Jest tak, ponieważ FEP jest tym, czym jest – zasadą. Nie można jej

[46] W tekstach tych można znaleźć odwołania do pracy Badcocka nt. *Evolutionary systems theory: A unifying meta-theory of psychological science* (2012), która poświęcona jest problemowi unifikacji w różnych subdyscyplinach psychologii. Badcock jednak nie wspomina nic w ramach wyznaczanych przez neomechanicyzm.

[47] Kluczowe dla tego rodzaju modelowania i wyjaśniania jest ograniczenie 3M (model – mechanizm – mapowanie): model zjawiska docelowego wyjaśnia to zjawisko w zakresie, w jakim (a) zmienne w modelu odpowiadają możliwym do zidentyfikowania komponentom, działaniom oraz własnościom organizacyjnym mechanizmu docelowego, który wytwarza, utrzymuje lub leży u podstaw tego zjawiska oraz (b) zależności (być może matematyczne) znajdujące się wśród tych (być może matematycznych) zmiennych w modelu odpowiadają relacjom przyczynowym pomiędzy komponentami mechanizmu docelowego (Kaplan, 2011, s. 347).

sfalsyfikować, podobnie jak zasady stacjonarnego działania Hamiltona[48]. Nie można jej również obalić. W rzeczywistości niewiele można w tej kwestii zrobić. Ewentualnie można zapytać tylko o to, czy mierzalne systemy są zgodne z tą zasadą.

Niefalsyfikowalność FEP można rozumieć na dwa sposoby: albo FEP wyjaśnia wszystko, zatem nie ma takiego organizmu, jaki można by zastosować w falsyfikacji tej zasady; albo przeciwnie – FEP nie wyjaśnia niczego. Można się tutaj powołać na rozważania Ludwiga Wittgensteina, który twierdził, że zdania niemające negacji są albo pozbawione sensu, albo są normami zastosowania wyrażeń. Negując zdanie „Systemy (żywe – MP), które nie minimalizują energii swobodnej, nie istnieją" (Friston, 2013c, s. 2), nie tyle mówimy coś na temat istnienia systemów, które nie realizują takiej czy innej funkcji, ile wypowiadamy pewną regułę konstytuującą nasz sposób myślenia i mówienia o organizmach żywych. W tym znaczeniu można bronić normatywnego wymiaru FEP, niemniej wydaje się, że nie o to Fristonowi chodzi[49]. Niektórzy badacze utrzymują, wbrew znanej tezie Karla Poppera, że falsyfikowalność nie jest ani warunkiem koniecznym, ani warunkiem wystarczalnym dla tego, aby dane wyjaśnienie było prawdziwe. W tym sensie FEP mogłaby pełnić funkcję zasady ograniczającej zakres możliwych wyjaśnień (por. Sims, 2017, s. 9). Kwestia ta, jak widać, nie jest przesądzona.

Friston sugeruje, że FEP można porównać do sformułowanej na gruncie fizyki zasady Hamiltona (Friston, Fortier, Friedman, 2018; por. Hohwy, 2020b[50]). Analogia ta jest nie tylko niejasna, ale także zwodnicza. Powszechnie uważa się, że wyjaśnienia w naukach fizycznych różnią się od wyjaśnień formułowanych na gruncie biologii (por. Braillard, Malaterre, 2015; Hopfield, 1994). Część badaczy, mówiąc o wyjaśnieniach, zwraca w tym kontekście uwagę na rolę tzw. kompromisów (*trade-off*) w modelowaniu (por. Levins, 1966). Zdaniem Evansa i współautorów (Evans i in., 2013) w naukach biologicznych występuje kompromis pomiędzy dążeniem do prostoty a ogólnością wyjaśniania. Kompromis ten nie występuje w naukach fizycznych. Oznacza to, że w naukach fizycznych im prostszy model, tym więcej wyjaśnia. W naukach biologicznych nie mamy jednak do czynienia z taką sytuacją. Zdaniem tych badaczy, im prostszy jest model biologiczny, tym bardziej formułowane na jego gruncie wyjaśnienia są ograniczone i lokalne. Przemysław Nowakowski (2018) uważa, że wykorzystywanie zbyt prostych mo-

[48] Zasada ta głosi, że rzeczywisty układ fizyczny porusza się po trajektorii, dla której działanie Hamiltona przyjmuje wartość stacjonarną (tj. minimum, maksimum lub punkt przegięcia), przy czym w obliczaniu działania rozważa się wszystkie możliwe trajektorie łączące zadany punkt początkowy i końcowy w zadanym czasie. Jeżeli punkty te leżą blisko siebie, to działanie ma minimum.

[49] Hohwy (2020b) twierdzi, że FEP jest teorią normatywną w tym sensie, że oferuje pojęcia, które służą do wyjaśnienia funkcjonowania organizmów żywych jako systemów samoorganizujących się. O FEP można zatem myśleć jako o teorii regulatywnej w sensie Kanta.

[50] Hohwy (2020b) podkreśla, że z powodu niemożliwości sfalsyfikowania FEP, zasadzie tej jest bliżej do zasady Hamiltona niż do teorii ewolucji przez naturalną selekcję.

deli (a taki wydaje się model Fristona) do wyjaśniania zjawisk biologicznych powoduje, że będą one dotyczyć albo bardzo lokalnych własności, tzn. nie będą odnosiły się do procesów i mechanizmów; albo doprowadzą do takiej sytuacji, że zjawiska biologiczne zostaną ujęte jako zwyczajne zjawiska fizyczne. Oznacza to, że (w najlepszym przypadku) wyjaśnienia oferowane przez FEP stanowią jakąś niebanalną fizyczną charakterystykę złożonych systemów fizycznych, do których również można zaliczyć organizmy żywe. Otwarte pozostaje pytanie dotyczące tego, czy takie wyjaśnienia będą cenne dla badaczy życia?

Sformułowane przeze mnie pytania i wątpliwości dotyczące FEP prowadzą do sceptycyzmu w kwestii możliwości eksplanacyjnych tej zasady oraz jej ostatecznego zastosowania do takich zagadnień jak omawiana tutaj kwestia ciągłości pomiędzy życiem a umysłem. Mateo Colombo i Cory Wright (2018), widząc trudności, na jakie napotyka FEP, zastanawiają się, z jakiego rodzaju zasadą mamy tutaj do czynienia i jak jej (zakładane przez Fristona i jego zwolenników) możliwości eksplanacyjne wiążą się z dwoma podstawowymi stanowiskami teoretycznymi w dzisiejszej biologii, tj. organicyzmem i mechanicyzmem. Analiza rozważań Colombo i Wrighta pozwoli mi ostatecznie odpowiedzieć na sformułowane przeze mnie na początku tego rozdziału pytanie o to, czy ujęcie architektoniczne (które wiążę z PP) musi być z konieczności rozpatrywane i wyjaśniane w kontekście ujęcia homeostatycznego (które wiążę z FEP).

Colombo i Wright twierdzą, że pomimo wielu różnych sformułowań tej zasady, należy zauważyć, że rozumowanie prowadzące do FEP ma formę argumentu transcendentalnego, którego wnioskiem jest twierdzenie, że FEP jest warunkiem możliwości istnienia systemów adaptacyjnych, w tym systemów przetwarzających informacje i niebiologicznych systemów, takich jak sieci społeczne i artefakty (Friston, 2009, s. 293). Friston stwierdza: „systemy, które nie minimalizują energii swobodnej, nie mogą istnieć" (2013c, s. 2). Argument ten, w sformułowaniu Colombo i Wrighta (2018, s. 4), ma następującą formę:

1. Jeżeli system Σ działa selektywnie na środowisko w celu uniknięcia przejść fazowych[51] i nie jest w stanie stałej równowagi z tym środowiskiem, to Σ zachowuje się adaptacyjnie.
2. System Σ zachowuje się adaptacyjnie tylko wtedy, gdy Σ zachowuje swoją fizyczną integralność poprzez utrzymywanie jego „charakterystycznych" zmiennych w granicach homeostazy, pomimo zmian obecnych w środowisku.
3. System Σ działa selektywnie na środowisko w celu uniknięcia przejść fazowych i nie znajduje się w stanie równowagi tylko wtedy, gdy Σ zachowuje swoją fizyczną integralność poprzez utrzymanie „charakterystycznych" zmiennych w granicach homeostazy, pomimo zmian środowiskowych.

[51] Por. przykład z uskrzydlonymi płatkami śniegu (podrozdział *Homeostaza w zasadzie energii swobodnej*).

4. System Σ utrzymuje swoją fizyczną integralność, zachowując swoje „charakterystyczne" zmienne w granicach homeostazy, pomimo zmian środowiskowych i tylko w przypadku, gdy Σ ogranicza entropię informacji (średnie zaskoczenie) swoich potencjalnych stanów zmysłowych.

5. Jeżeli system Σ minimalizuje energię swobodną swoich możliwych stanów zmysłowych, to Σ zmniejsza entropię tych stanów.

6. Każdy system Σ, który zmniejsza entropię swoich możliwych stanów zmysłowych, utrzymuje swoją integralność fizyczną, zachowując jednocześnie „charakterystyczne" zmienne w granicach homeostazy, pomimo zmian środowiskowych.

7. Każdy system Σ, który minimalizuje energię swobodną swoich możliwych stanów zmysłowych, utrzymuje swoją integralność fizyczną poprzez zachowanie swoich „charakterystycznych" zmiennych w granicach homeostazy, pomimo zmian środowiskowych[52].

Rozumowanie to prowadzi do stwierdzenia, że FEP jest pierwszą zasadą *a priori*, konieczną dla pojawienia się i (zrozumienia) funkcjonowania systemów adaptacyjnych, w tym także tych systemów, należy to dodać, które realizują mechanizmy PP. Punkt 1 określa warunek termodynamiczny; punkt 2 warunek homeostatyczny. W punkcie 3 wiążą się fizyczne i biologiczne określenia systemu adaptacyjnego. Punkt 4 wiąże homeostazę z wymiarem informacyjno-teoretycznym, zaś 5 orzeka, że minimalizacja energii swobodnej jest warunkiem wystarczającym ograniczania entropii stanów zmysłowych danego systemu. W punkcie 6 stwierdza się warunek konieczny tego, aby system utrzymał swoją integralność fizyczną; zaś punkt 7 wyraża FEP[53].

Powstaje teraz pytanie o to, czy takie rozumowanie może być zasadnie wykorzystane w wyjaśnieniach formułowanych przez badaczy. Aby odpowiedzieć na to pytanie, Colombo i Wright proponują, aby odnieść FEP do dwóch najważniejszych nurtów teoretycznych we współczesnej biologii, tj. organicyzmu oraz mechanicyzmu. Przypomnijmy: w proponowanej tutaj przeze mnie wykładni organicyzm należy wiązać z non-kognitywizmem i niektórymi radykalnymi ujęciami PP, mechanicyzm zaś z kognitywizmem i konserwatywnym PP. Wcześniej omówiłem wiele możliwych zarzutów wobec zastosowania FEP zarówno w jednym, jak i drugim podejściu. Obecnie przechodzę do ostatniej części analiz poświęconych możliwościom eksplanacyjnym FEP. Krytyka przeprowadzona przez Colombo i Wrighta

[52] Można jednak twierdzić, że sugerowałem w rozdziale *Pojęcie przetwarzania predykcyjnego*, że tego typu argumentacja nie jest transcendentalna, ale ma charakter wnioskowania abdukcyjnego. Jeżeliby tak było, to rozumowanie Fristona ma jeszcze słabszą siłę dowodową.

[53] Istnieją jednak argumenty, które wykazują słabość skonstruowanego przez Colombo i Wrighta argumentu transcendentalnego na rzecz FEP. Trzeba bowiem zauważyć, że entropia informacji może osiągnąć stałą wartość, a jednak energia swobodna może nadal wzrastać. Oznacza to, że minimalizowanie energii swobodnej jest wystarczającym, ale niekoniecznym warunkiem do umieszczenia górnej granicy na entropii. Nie może być w takim razie warunkiem transcendentalnym. Problem zatem dotyczy przejścia pomiędzy punktami 5 a 6 argumentu. Inną sprawą jest fakt, że nie jest jasne przejście od energii w sensie termodynamicznym do energii w sensie informacyjno-teoretycznym. Dziękuje za te uwagi Tomaszowi Korbakowi i Marcinowi Miłkowskiemu.

uzupełni bronione tutaj stanowisko, odmawiające zasadzie sformułowanej przez Fristona wartości eksplanacyjnej, którą niektórzy badacze jej przypisują.

Zdaniem tych badaczy FEP nie jest zgodna ani z organicyzmem, ani mechanicyzmem. Nie wchodząc w szczegóły drobiazgowej analizy dokonanej przez Colombo i Wrighta, należy powiedzieć, że argument transcendentalny, uzasadniający FEP jako pierwszą zasadę w naukach o życiu, nie może zostać w całości zastosowany w żadnym z dominujących poglądów na wyjaśnienie życia.

Wobec zastosowania FEP w organicyzmie Colombo i Wright formułują m.in. następujące zarzuty:

1. FEP zakłada m.in., że wszelkie organizmy są ergodyczne. Nie zgadzają się z tym poglądem organicyści, ponieważ ich zdaniem cecha ergodyczności nie jest istotna dla żywych organizmów (por. Longo, Montévil, Kauffman, 2012).

2. Według FEP dynamikę adaptacyjną organizmów można wyjaśnić jako zmianę ich przestrzeni fazowej. Organicyści uważają, że dynamiki adaptacyjnej nie można odwzorowywać w ten sposób. Zmiany w przestrzeni fazowej są historyczne i nabyte ewolucyjnie, zatem nie można traktować ich jako specyficznych dla danego osobnika, a jedynie jako „gatunkowe".

3. Zwolennicy FEP przyjmują w biologii stanowisko adaptacyjne i selekcyjne przy założeniu, że „selekcja wyjaśnia sposób, w jaki powstają systemy biologiczne" (Friston, Stephan, 2007, s. 423). Organicyści odrzucają taką perspektywę, twierdząc, że mechanizmy adaptacji i selekcji nie są wystarczające do wyjaśnienia autonomii żywych organizmów oraz ich procesów regulacyjnych w odniesieniu do środowiska. Przykładowo Giuseppe Longo i Maël Montévil (2013) twierdzą, że zachowania adaptacyjne organizmów powinny być reprezentowane jako ciągłe przejście z jednej fazy do drugiej, w wyniku czego organizmy są wciąż odtwarzane poprzez nieustanne sprzężenie ze środowiskiem. Podstawową ideą jest to, że z jednej strony to przodkowie stale współtworzą i przekształcają organizmy, z drugiej zaś, że na to przekształcenie wpływa ciągła interakcja z ekosystemem.

4. Zwolennicy FEP utrzymują, że ewolucja, a szczególnie procesy selekcji naturalnej, są także procesami minimalizacji energii swobodnej. Dla organicystów takie ujęcie przesłania jedynie historyczne oraz ekologiczne ugruntowanie adaptacji biologicznej i nie wnosi nic do zrozumienia natury życia.

Głównym powodem odrzucenia wykorzystania FEP w mechanicyzmie[54] jest fakt, że analizy funkcjonalne proponowane przez zwolenników FEP nie mają wystarczającej mocy eksplanacyjnej, ponieważ można je potraktować jedynie jako szkice mechanizmów (por. Piccinini, Craver, 2011). Szkic mechanizmu jest niekompletną reprezentacją danego mechanizmu, w której pomijane są pewne aspekty wyjaśnianego zjawiska (np. niektóre komponenty, operacje lub część orga-

[54] Do kwestii wyjaśnień mechanistycznych wracam w rozdziałach *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego* i *Mechanizmy normatywne a działania w przetwarzaniu predykcyjnym*.

nizacji mechanizmu). Satysfakcjonujące wyjaśnienie badanego zjawiska opiera się dopiero na sformułowaniu schematu mechanizmu, który odnosi się do faktycznych komponentów mechanizmu oraz jego organizacji. Schemat mechanizmu jest faktycznie kompletnym modelem danego zjawiska (Craver, 2007, s. 113–114). Analizy funkcjonalne w FEP są idealizacjami, abstrahującymi od szczegółów biologicznych i fizycznych, które dotyczą złożoności strukturalnej rzeczywistych obiektów. Analizy pozbawione odwołania do tych parametrów nie mają mocy eksplanacyjnej i są jedynie pewnymi abstrakcyjnymi idealizacjami. Oprócz tego należy powiedzieć, że niejasny jest status epistemiczny tej zasady wobec wyjaśnień mechanistycznych. Zależnie od kontekstu FEP określana jest jako imperatyw, tautologia, prawo życia, paradygmat, aksjomat, wyjaśnienie unifikujące czy pierwsza zasada *a priori*. Taka rozpiętość rozumienia nie sprzyja zupełnie określeniu statusu tej zasady. Zdaniem Colombo i Wrighta FEP, w najlepszym wypadku, można traktować jako jedno z istotnych ograniczeń nakładanych na wyjaśnienia mechanistyczne w naukach o życiu, o ile badacze będą chcieli określić, które z funkcji energetycznych są istotne dla utrzymania homeostazy. Autorzy podkreślają także, że olbrzymi zakres zastosowania FEP pokazuje, że w naukach biologicznych jest miejsce na dedukcyjną systematyczność i fizyczne podstawy. Stanowczo jednak odrzucają przekonanie Hohwy'ego (por. 2014, 2015), jakoby FEP była zgodna i komplementarna z mechanicyzmem (Colombo, Wright, 2018, s. 23).

3.6. Zastosowanie zasady energii swobodnej w przetwarzaniu predykcyjnym

Po przeprowadzonych w tym rozdziale szerokich analizach można przejść do odpowiedzi na pytanie wyjściowe, które dotyczyło tego, czy ujęcie architektoniczne (wiązane z PP) musi lub powinno być oparte na ujęciu homeostatycznym albo rozważane w odniesieniu do tego ujęcia. Moja odpowiedź jest negatywna. FEP nie jest zasadą, która jest konieczna dla wyjaśnienia mechanizmu PP. Nie jest również ramą teoretyczną, która jest konieczna, aby zrozumieć PP. Zachowuję zatem sceptycyzm wobec stanowisk zarówno kognitywistycznego, jak i nonkognitywistycznego w kwestii zastosowania FEP w PP. Moim zadaniem nie jest tutaj uznanie bądź odrzucenie wszelkich pretensji eksplanacyjnych zwolenników tej zasady, niemniej uważam, że z powodów, które podałem, można analizować i prowadzić badania nad PP, zupełnie abstrahując od wymiaru biologicznego utożsamianego z procesami minimalizacji energii swobodnej oraz redukcji entropii. Idę tutaj za sugestią Orlandi (2017), która proponuje, aby odróżnić PP ujęte szeroko (poprzez odniesienie do FEP) od PP ujętego wąsko i rozumianego jako teoria mechanizmów percepcyjnych. Szerokie PP z jednej strony ma pozwolić na wyjaśnienie olbrzymiej klasy zjawisk (np. wyższe

procesy poznawcze, pamięć, działanie, myślenie; por. Clark, 2013b; Hohwy, 2013, 2015; Seth, 2015), z drugiej zaś staje się podatne na słabości podzielane przez FEP, a także zarzuty formułowane przez krytyków dotyczące albo architektury modelu generatywnego, albo możliwości wyjaśnienia poszczególnych mechanizmów (por. Klein, 2018; Williams, 2018). Wąsko ujęte PP odnosi się jedynie do wyjaśnienia mechanizmów predykcyjnych, które są konstytutywne i przyczynowo efektywne dla określonych procesów percepcyjno-motorycznych. Ta węższa perspektywa retorycznie wzbudza mniejszy podziw, niemniej, co będę dalej uzasadniał, oferuje satysfakcjonujące wyjaśnienie wielu mechanizmów, które są odpowiedzialne za realizację szeregu funkcji poznawczych. Uwolnione PP od „gorsetu" FEP pozwala na sformułowanie nietrywialnego wyjaśnienia normatywnych własności mechanizmów predykcyjnych, co w dalszej perspektywie pozwala na rozwiązanie wielu problemów (np. „problem ciemnego pokoju"), z którymi szeroko interpretowane PP nie radzi sobie w sposób satysfakcjonujący[55].

Odrzucenie pretensji eksplanacyjnych FEP nie oznacza jednocześnie odrzucenia wielu interesujących rozwiązań sformułowanych przez Fristona i jego współpracowników. Nie odrzucam zatem badań nad FEP i jego zastosowaniem w dziedzinie filozofii kognitywistyki, ile odrzucam możliwość wyjaśnienia PP i w efekcie ludzkiego poznania poprzez odwołanie się do FEP jako pierwszej i unifikującej zasady *a priori* lub mówiąc innym językiem: nie uznaję transcendentalnego argumentu, który dowodzi, że procesy minimalizacji energii swobodnej są warunkiem koniecznym pojawienia się procesów poznawczych.

W rozdziale *Pojęcie przetwarzania predykcyjnego* wskazałem na istnienie wielu możliwych algorytmów dotyczących tego, w jaki sposób realizowane są mechanizmy PP (por. podrozdział *Możliwe algorytmy przetwarzania predykcyjnego i filozoficzne inspiracje*). Przypomnijmy: analizy na poziomie algorytmicznym są kluczowe dla badaczy, ponieważ pozwalają na wyjaśnienie modelowanych procesów poznawczych (Marr, 1982). Z wielu względów model PP oparty na FEP wydawał się najlepszym kandydatem wyjaśniającym poziom algorytmiczny realizacji mechanizmów PP. Analizy w tym rozdziale wykazały jednak, że FEP nie jest odpowiednią ramą wyjaśniającą dla PP. Ogólnie mówiąc, FEP nie oferuje satysfakcjonującego wyjaśnienia zjawisk, które obiecuje wyjaśnić. Dodatkowo nie jest jasny jej status metodologiczny oraz perspektywa badawcza. Trudno jest jednoznacznie wyjaśnić, na czym ma polegać aprioryczny i fundujący charakter tej zasady zarówno w odniesieniu do perspektyw reprezentacjonistycznej i antyreprezentacjonistycznej, jak i organistycznej oraz mechanistycznej.

Jak wcześniej wspomniałem, odrzucenie pretensji eksplanacyjnych FEP nie jest jednocześnie odrzuceniem całej teorii, która towarzyszy temu zagadnieniu. Twierdzę, że badania Fristona nad FEP stanowią ważną heurystykę badawczą

[55] Na ten temat por. rozdział *Mechanizmy normatywne a działania w przetwarzaniu predykcyjnym*.

dla PP[56]. Niech przykładem takiego heurystycznego wpływu będzie normatywny wymiar FEP. Friston i jego współpracownicy niejednokrotnie określają FEP jako tzw. zasadę normatywną (por. Allen, Friston, 2018, s. 2476; Friston, Fortier, Friedman, 2018, s. 21). Friston wprost nie definiuje, co rozumie przez normatywność, niemniej można ją wiązać z wyborem optymalnego działania w warunkach niepewności. Optymalne decyzje minimalizują różnicę pomiędzy entropią (rozumianą tutaj jako niepewność oczekiwań) a przewidywaniami (por. Schwartenbeck i in., 2013). System poznawczy nie tylko minimalizuje tę różnicę, ale przede wszystkim *powinien* ją minimalizować, aby skutecznie działać w swoim środowisku. Ten normatywny wymiar FEP jest niezwykle istotny ze względu na realizowane w tej pracy cele poznawcze. Niemniej normatywność FEP wymaga pełnej eksplikacji. Jej opis zaproponowany przez Fristona jest wieloznaczny i niejasny. Po pierwsze nie wiadomo, w jakim sensie FEP jest zasadą normatywną; po drugie należy wykazać, dlaczego mechanizmy minimalizacji energii swobodnej są normatywne z punktu widzenia organizmu i co to właściwie znaczy, że są normatywne; po trzecie zaś, czy mówiąc o normatywności w tym kontekście, ma się na myśli funkcje normatywne, działania, czy jeszcze coś innego? Odpowiedzią na te pytania zajmę się w rozdziale *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego*.

Podsumowanie

W niniejszym rozdziale zastanawiałem się, czy PP musi być, jak twierdzą niektórzy badacze, wyjaśniane w odniesieniu do fristonowskiej FEP. Zwróciłem uwagę, odnosząc się do sporu dotyczącego ciągłości pomiędzy procesami konstytuującymi życie a procesami konstytuującymi poznanie, że istnieje wiele różnych, często wykluczających się, wykładni tej zasady. Niektórzy wiążą ją z informacją syntaktyczną, inni zaś z informacją semantyczną i wyjaśnieniami opartymi na reprezentacjach. Niemniej, niezależnie od interpretacji FEP, niejasnym pozostaje jej rozumienie jako pierwszej i koniecznej zasady *a priori*, a także nie wiadomo, w jaki sposób FEP wyjaśnia zjawiska, które zdaniem jej zwolenników ma wyjaśnić. Analizując wiele argumentów za i przeciwko zastosowaniu FEP w PP, stwierdziłem, że FEP nie stanowi satysfakcjonującej ramy eksplanacyjnej dla PP oraz że istnieje wiele innych algorytmów wskazujących na możliwe wyjaśnienie mechanizmów predykcyjnych. Argumentowałem za tym, że FEP stanowi istotną i niebanalną heurystykę badawczą dla analiz normatywności mechanizmów predykcyjnych. Znaczenie tej heurystyki zostanie omówione w następnym rozdziale.

[56] Przez heurystykę rozumiem tutaj zdolność do rozpoznawania takich danych, jakie następnie mogą być skutecznie wykorzystywane do rozwiązywania określonych problemów. Innymi słowy, heurystyka jest poszukiwaniem wcześniej niezdefiniowanych wzorów (działania, myślenia, wyjaśniania itd.). Por. Turing, 1968.

Rozdział iv
Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego

Wstęp

Celem niniejszego rozdziału jest eksplikacja pojęcia funkcji i mechanizmów normatywnych w kontekście PP. Rozpocznę od wyjaśnienia, w jakim sensie FEP jest normatywna i czy jej normatywność pozwala uzasadnić normatywność mechanizmów predykcyjnych. Następnie zrekonstruuję i krytycznie przeanalizuję wybrane koncepcje funkcji biologicznych. Będę twierdził, że funkcje normatywne w proponowanym przeze mnie ujęciu powinny spełniać dwa wymogi: eksplanacyjny i normatywny. Wykażę, że żadne z omawianych stanowisk nie spełnia tych wymogów w sposób zadowalający. Będę bronił zmodyfikowanej wersji ujęcia Bickharda, twierdząc, że specjalnym przypadkiem funkcji normatywnych są funkcje reprezentacyjne i predykcyjne.

W dalszej części rozdziału sformułuję model normatywności dla PP. Stwierdzę, że określona funkcja jest normatywna wtedy, gdy jednocześnie (1) pozwala wyjaśnić dysfunkcję danego mechanizmu; (2) przyczynia się do utrzymania stabilności organizmu w ten sposób, że kształtuje i ogranicza możliwe relacje, procesy i zachowania danego systemu; oraz (3) – w odniesieniu do funkcji reprezentowania i funkcji predykcyjnej – gdy można za jej pomocą wyjaśnić możliwość przypisywania wartości logicznych określonym reprezentacjom/predykcjom. Mechanizm zaś jest normatywny wtedy, gdy realizuje określone funkcje normatywne i gdy jest możliwe, aby dany system był mechanizmem dla jakiejś aktywności, pomimo faktu, że (np. w danym momencie) nie może jej wykonywać. Takie ujęcie jednak, co wykażę, jest eksplanacyjnie słabe. Powiążę je ze stanowiskiem preskryptywizmu ontycznego. Zaproponuję ontyczną interpretację pojęcia mechanizmu normatywnego. Zgodnie z nią, jeżeli orzeka się, że dany mechanizm albo funkcja jest normatywny, to twierdzi się, że spełnia on takie, a nie inne funkcje przyczynowe. W odniesieniu do normatywnych własności mechanizmów i funkcji predykcyjnych oznacza to, że są one przyczynami określonych działań danego organizmu w środowisku. Innymi słowy, normatywność jest takim predykatem, za pomocą którego możemy wyjaśnić pewne zjawiska (tj. działania i zachowania) w kategoriach mechanizmów i funkcji, które są ich przyczynami.

Na koniec wykażę, że właściwą ramą eksplanacyjną dla tak rozumianych funkcji i mechanizmów predykcyjnych jest neomechanistyczna koncepcja wyjaśnień naukowych. W ten sposób uzasadnię, że (1) teoretycznie PP ma zastosowanie na wszystkich trzech poziomach analizy w rozumieniu Marra; (2) stanowisko preskryptywizmu ontycznego (tj. realizmu) w odniesieniu do PP jest zasadne z powodu ontycznego charakteru wyjaśnień formułowanych w ramach tego podejścia; oraz (3) bayesowski model PP jest nietrywialną perspektywą badawczą ze względu na normatywny charakter bayesowskich mechanizmów predykcyjnych.

4.1. Uwagi wprowadzające

Celem analiz przeprowadzonych w ostatnim rozdziale była odpowiedź na pytanie dotyczące relacji pomiędzy PP a FEP. Przypomnijmy: część badaczy uważa, że FEP stanowi istotną ramę eksplanacyjną dla PP i pozwala wyjaśnić, dlaczego systemy poznawcze minimalizują swoje błędy predykcyjne. W myśl tego ujęcia, mózg, rozumiany jako model generatywny, jest specjalnym przypadkiem systemu minimalizującego swoją energię swobodną (w sensie teoretyczno-informacyjnym), FEP zaś jest tą zasadą, która wyjaśnia, dlaczego w sposób konieczny ergodyczne układy żywe dążą, ale także *muszą* dążyć, do redukcji swojego zaskoczenia, czyli różnicy pomiędzy tym, w jaki sposób świat jest modelowany przez organizm, a tym, jaki świat *jest*. Minimalizacja energii swobodnej prowadzić ma do osiągnięcia stanu homeostazy (równowagi energetycznej – w sensie teoretyczno-informacyjnym – ze środowiskiem) rozumianego jako stan relatywnie niskiej entropii. Wykazałem, że FEP nie jest odpowiednią ramą eksplanacyjną dla PP, ponieważ faktycznie niewiele mówi o naturze systemów poznawczych, a przede wszystkim nie odnosi się do żadnych określonych mechanizmów. Innymi słowy, nie wyjaśnia tego, co obiecuje wyjaśnić. Moja konkluzja jednak nie była zupełnie negatywna. Postawiłem hipotezę, że FEP oraz powiązane z nią badania Fristona i jego współpracowników stanowią istotną i nietrywialną heurystykę badawczą dla analiz z zakresu PP. Owa heurystyka bezpośrednio wiąże się z normatywnym wymiarem FEP i tzw. wnioskowania aktywnego. Teraz przyjdzie mi uzasadnić tę hipotezę.

Niniejsze badania są szczególnie istotne ze względu na realizację celu postawionego przed tą pracą, czyli wykazania, że mechanizmy predykcyjne są normatywne. Moje analizy rozpocznę od pytania dotyczącego normatywności FEP.

Friston i współpracownicy wielokrotnie podkreślają, że FEP jest zasadą normatywną (por. Adams, Brown, Friston, 2014; Allen, Friston, 2018; Friston, 2013; Friston, Fortier, Friedman, 2018). Wprost jednak nie wyjaśniają, na czym jej normatywność polega. Twierdzą, że FEP jest normatywny, ponieważ (1) mówi, że pewne rzeczy są z nią zgodne, a inne nie (Friston, Fortier, Friedman, 2018, s. 21);

(2) dostarcza normatywny opis odnoszący się do tego, że mózg, aby zachować swoją własną integralność, *musi* angażować się w minimalizowanie energii swobodnej, *resp*. błędów predykcyjnych (Allen, Friston, 2018, s. 2472); (3) wywodzi się *a priori* z definicji organizmów żywych, która głosi, że życie polega na unikaniu entropii (Allen, Friston, 2018, s. 2473); (4) odnosi się do normatywności wariacyjnej energii swobodnej, ponieważ pozwala ona zdefiniować działania realizowane przez dany system na poziomie obliczeniowym (Friston, Fortier, Friedman, 2018, s. 23; por. Marr, 1982). O normatywności FEP można również mówić na metapoziomie, tzn. odnosząc ją, nie tyle do charakterystyki jej treści, ile do jej statusu eksplanacyjnego i teoretycznego. W tym aspekcie należy powiedzieć, że FEP jest normatywna, ponieważ pozwala na syntezę biologii i teorii informacji (Allen, Friston, 2018, s. 2476).

Na poziomie przedmiotowym FEP jest zatem normatywna, czyli ze względu na swoją treść, oraz na poziomie metaprzedmiotowym, czyli ze względu na swoją funkcję. Można te dwa aspekty tej zasady rozważać w łączności, tzn. utrzymywać, że np. metanormatywność wywodzi się z normatywności przedmiotowej; albo niezależnie od siebie. Sądzę jednak, że na pytanie o normatywność FEP jako metazasady już pośrednio odpowiedziałem. Przytoczone uwagi Fristona dotyczą unifikującego charakteru FEP jako zasady syntetyzującej różne dziedziny wiedzy. Wydaje się bowiem, że FEP ma nie tylko umożliwić powiązanie biologii i teorii informacji, ale także fizyki teoretycznej, neuronauk, niektórych działów psychologii, psychiatrii czy nawet socjologii, bowiem zasada ta ma mieć zastosowanie „od organizmów jednokomórkowych do sieci społecznych" (Friston, 2009, s. 293). W rozdziale *Zasada energii swobodnej w przetwarzaniu predykcyjnym* wykazałem, że te unifikujące zapędy FEP, oparte na jej pretensjach eksplanacyjnych, są przedwczesne. Głównym powodem jest jej słabość eksplanacyjna: nie da się mówić o unifikacji różnych dziedzin wiedzy na podstawie zasady, która nie odwołuje się do żadnych faktycznych mechanizmów, a jedynie do pewnych wyidealizowanych uogólnień. Pozostaje zatem analiza FEP jako zasady normatywnej w sensie przedmiotowym. Zanim przejdę do tej analizy, konieczne jest omówienie samego pojęcia normatywności i jego możliwych rozumień.

Otwarty pozostaje również problem tego, co właściwie stanowi źródło normatywności FEP oraz czy normatywność, o której tutaj mowa, sprowadzalna jest do jakichś faktów, czy może jest ona pierwotna względem faktów. Jeżeli chodzi o to ostatnie pytanie, to Friston kategorycznie odrzuca tłumaczenie normatywności FEP poprzez odniesienie jej do jakichś uprzednich stanów rzeczy albo przedmiotów (por. Allen, Friston, s. 2018; Friston, Fortier, Friedman, 2018). To jednak rodzi kolejny problem: skoro normatywności FEP nie można sprowadzić czy wyjaśnić poprzez odniesienie do jakichś faktycznych własności, to na jakich własnościach opiera się ta zasada? Musi opierać się na jakichś własnościach, mechanizmach czy procesach, które są normatywne *per se*. Co do tego nie ma wątpliwości, bowiem

sam Friston nieustannie podkreśla, że FEP wywodzi się z pewnych cech organizmów żywych, a konkretnie z ich tendencji do unikania entropii, czyli z potrzeby utrzymania homeostazy, o czym już wspomniałem powyżej. Takie ujęcie jednak prowadzi do kolejnych trudności, ponieważ nie wiadomo, czym jest owo „dążenie" do homeostazy. Czy jest to pewna funkcja organizmów żywych, czy raczej jakiś proces? A może po prostu ich własność?

Podsumowując powyższy wątek, stwierdzam, że nie można przypisać normatywności FEP, ponieważ jej normatywność musi wywodzić się z czegoś bardziej podstawowego. Ową normatywność FEP jedynie artykułuje. Sytuacja jest podobna do tej, którą znamy z klasycznej etyki normatywnej. Etyka ta jest normatywna, ponieważ formułuje ogólne normy postępowania. Te normy, rozumiane jako pewne zasady, są jednak wyrazem określonych prawidłowości, z którymi podmioty moralne mają do czynienia w świecie. Tym, co odróżnia etykę od prawa, jest wywodzenie określonych norm z bardziej pierwotnych normatywnie wartości, powinności czy jakichś przedmiotów bądź osób[1]. Korzystając z tej analogii, trzeba powiedzieć, że FEP jest normatywna, ponieważ odnosi się do *czegoś*, co jest normatywne w charakterystyce organizmów żywych. Wydaje się, że owym *tym*, co jest normatywne, jest, jak już wspomniałem, dążenie do homeostazy. Tym problemem zajmę się w moich następnych analizach.

4.2. Normatywność homeostazy

Pojęcie homeostazy w koncepcji Fristona jest ściśle powiązane z koniecznością minimalizacji energii swobodnej (por. Friston, 2013b, s. 38). Pojęcie to wprowadził w 1932 roku Walter Cannon. Tradycyjnie homeostaza oznacza zdolność organizmu do samoregulacji, rozumianej jako utrzymywanie stałości parametrów systemu w obliczu dynamicznie zmieniającego się środowiska. Homeostaza obejmuje zarówno stabilność, jak i zmianę, jednak pomiędzy funkcjami pełniącymi te własności nie zachodzi relacja symetryczna. Mechanizm jest mechanizmem homeostatycznym tylko wtedy, gdy udział w przetrwaniu uzyskany przez zmienny zbiór własności przechodzi w zbiór innych własności (Godfrey-Smith, 1996, s. 77). Zdaniem Ashby'ego „przyjęty tutaj punkt widzenia jest taki, że stałość podstawowych zmiennych jest zasadniczo ważna i że aktywność innych zmiennych jest ważna tylko o tyle, o ile przyczynia się do utrzymania tej stałości" (Ashby, 1963, s. 67). Istnieje wiele mechanizmów utrzymujących homeostazę, zarówno fizjologicznych, jak i behawioralnych. Friston jednak homeostazę rozumie przede wszystkim w terminach teoretyczno-informacyjnych, co czyni jego propozycję kontrowersyjną, ponieważ

[1] Dla przeprowadzanych tutaj rozważań nie jest istotne, czy takie postępowanie jest zasadne oraz czy istnieją jakieś własności świata lub przedmioty, które można określić mianem „normatywnych" albo „moralnych".

– o czym już wspominałem w poprzednim rozdziale – nie jest jasne, o jaki rodzaj informacji chodzi: czy o semantyczną, czy o syntaktyczną. Nie jest również oczywisty charakter relacji zachodzącej pomiędzy homeostazą rozumianą jako kontrola fizjologiczno-behawioralna a homeostazą w sensie Fristona.

Oczywistą inspiracją Fristona jest model homeostatu zaproponowany przez Ashby'ego (1963). Homeostat to układ o ujemnym sprzężeniu zwrotnym, który ma zdolność samoregulacji. Przeciwdziała on działaniu czynników zewnętrznych, które mogą wpłynąć na ten układ, zmieniając go. Za sprawą cybernetyki homeostaza została, w pewnym sensie, oderwana od swojego biologicznego podłoża. W tym ujęciu polega na takim teleologicznie konstytuowanym przeciwdziałaniu przepływowi informacji, jaki zmniejsza możliwość oddziaływania danego układu na jego otoczenie. W ten sposób pojęcia samoregulacji i informacji wiążą się z utrzymywaniem przez system życia (Ashby, 1963, s. 274). Zdaniem Ashby'ego cała aktywność organizmów skierowana na zewnątrz ma charakter homeostatyczny i teleologiczny. W jakim sensie teologiczny? Arturo Rosenblueth, Norbert Wiener i Julian Bigelow (1943) argumentowali, że pojęcie celowości ma istotne zastosowanie eksplanacyjne w nauce, zarówno w odniesieniu do wyjaśnienia natury maszyn, jak i świadomego projektowania oraz planowania. Ich zdaniem pojęcie to również nadaje się do opisu systemów żywych, a przede wszystkim mechanizmów homeostatycznych. Mówiąc o celowości danego systemu, orzeka się, że dany system jest skierowany na cel, czyli że zachowuje się w pewien określony sposób. Zachowanie to musi być jednak wyjaśnione poprzez odniesienie do jakiegoś wewnętrznego mechanizmu. Mechanizm ten w cybernetyce należy wiązać z ujemnym sprzężeniem zwrotnym (por. Garson, 2016, s. 25). Rosenblueth, Wiener i Bigelow (1943, s. 19) wprost twierdzą, że „wszystkie celowe zachowania mają być rozważane w odniesieniu do ujemnego sprzężenia zwrotnego". Wprost określają działania oparte na tym mechanizmie jako celowe. System sprzężenia zwrotnego jest specjalnym systemem typu wejście – wyjście. Pobiera on dane ze swojego środowiska, a następnie je przetwarza, w wyniku czego wytwarza pewne efekty na wyjściu, np. określone zachowania. System oparty na ujemnym sprzężeniu zwrotnym pobiera pewną informację „na wejściu", po czym wykorzystuje ją do określonego działania „na wyjściu". Dobrym przykładem takiego systemu jest termostat. Termostat wykrywa temperaturę pomieszczenia (na swoim wejściu), następnie zaś wykorzystuje informacje umożliwiające wytworzenie określonej mocy wyjściowej (np. włączenie pieca). W przypadku gdy temperatura pomieszczenia wzrasta (w wyniku działania termostatu), termostat podejmuje inne działanie (np. wyłącza piec). Z tego wynika, że system sprzężenia zwrotnego musi być wyposażony w czujnik, który monitoruje niektóre cechy środowiska, i efektor, który generuje zachowanie. W szczególności efektor musi być w stanie wytworzyć zachowanie, które modyfikuje środowisko w taki sposób, że dany czujnik może go wykryć. Należy wyróżnić dwa rodzaje

sprzężeń zwrotnych: dodatnie i ujemne. W pierwszym przypadku część energii pobranej przez system na wyjściu powraca do niego i wytwarza więcej takich samych zachowań. Następuje wzmocnienie danego zachowania. W przypadku ujemnego sprzężenia zwrotnego część energii pobranej na wyjściu powraca do systemu i wytwarza odwrotny rodzaj zachowania. Prowadzi to w efekcie do stłumienia tego zachowania. To ujemne sprzężenie zwrotne jest istotne, ponieważ na jego zasadzie działają zachowania samoregulacji obecne w przyrodzie (Garson, 2016, s. 26). Innymi słowy, homeostaza jest opisywalna w kategoriach ujemnego sprzężenia zwrotnego.

W tym kontekście pojawia się kluczowe dla Ashby'ego pojęcie „utrzymywania się przy życiu", które należy rozumieć jako synonim „stabilności", pewnego zbioru[2]. Zbiór ten jest koniunkcją różnych stanów, w których dany układ mógł i może się znajdować w określonych okolicznościach. Rozważania te prowadzą Ashby'ego do stwierdzenia, że tym, co utrzymuje się przy życiu z upływem czasu, nie jest indywidualny organizm, lecz pewne zbiory genów, które prowadzą do wytworzenia określonego indywiduum. Te geny, które nie prowadzą do rozwoju danego organizmu, w trakcie ewolucji są eliminowane (Ashby, 1963, s. 276–277). Przepływ informacji pomiędzy organizmem a otoczeniem dotyczy pierwotnie wymiany informacyjnej na poziomie genów. Zatem, mówiąc językiem cybernetyki, to negatywne sprzężenie zwrotne zachodzi głównie na poziomie genetycznym. Wymiana informacyjna na wyższych poziomach organizacji systemu opiera się na wymianie informacji genetycznej. Takie ujęcie rodzi dwa problemy: pierwszy dotyczy natury informacji, o czym już pisałem wcześniej, drugi zaś powiązania perspektywy Ashby'ego i Fristona. Problematyczne jest bowiem rozumienie homeostazy u obu tych autorów. W ujęciu Fristona dotyczy ona przede wszystkim indywiduum i jest obecna na różnych poziomach jego organizacji. Jest zatem realizowana na poziomie fenotypu. U Ashby'ego zaś odnosi się do poziomu gatunkowego, jest bowiem homeostazą realizowaną na poziomie genotypu. W jednym i drugim przypadku istotny jest teleologiczny charakter mechanizmów homeostatycznych. Oznacza to, że np. mechanizm termoregulacji jest ukierunkowany na utrzymanie homeostazy. Ma ona zatem charakter celowy (por. Craik, 1966, s. 12). W ujęciu Ashby'ego będzie to mechanizm genetycznie nabyty i warunkowany na poziomie gatunkowym, zaś w przypadku Fristona będzie to mechanizm realizowany przez określone indywiduum, niezależnie od jego historii ewolucyjnej[3]. Friston wprost nie odwołuje się do pojęcia ujemnego sprzężenia zwrotnego. Jedynym wyjątkiem

[2] Craik powiada, że żywe organizmy umierają, jeżeli nie utrzymują dynamicznej równowagi ze środowiskiem. „Jeżeli zaś ją utrzymują, to wykazują pewien stopień spontaniczności, zmienności i celowości, odpowiadając na to, co nieznane w świecie nieożywionym" (1966, s. 12).

[3] Nie należy tego rozumieć w tym sensie, że historia ewolucyjna danego narządu czy mechanizmu jest nieistotna. Chodzi raczej o to, że wyjaśnienie danego mechanizmu nie musi mieć charakteru ewolucyjnego. Wystarczy sam opis funkcjonalny. Więcej na ten temat por. podrozdział *Funkcje w ujęciu Wrighta i Cumminsa*.

jest tekst poświęcony nastrojom (por. Clark, Watson, Friston, 2018; por. także Rao, Ballard, 1999), niemniej jego analizy homeostazy są rozumiane podobnie. Dodatkowym argumentem za cybernetycznym wymiarem jego rozumienia homeostazy jest odwoływanie się do mechanizmu minimalizacji błędu predykcyjnego, który także był istotny dla klasycznej cybernetyki (por. analiza działania torpedy jako ukierunkowanej na pewien cel – Rosenblueth, Wiener, Bigelow, 1943, s. 20).

To, co jest istotne, to fakt, że mechanizmy homeostazy w rozumieniu Fristona są mechanizmami teleologicznymi. To pozwala o nich myśleć w kategoriach normatywnych. Mianowicie: mechanizm, który działa w sposób celowy (np. wpływa na utrzymanie homeostazy), działa poprawnie. Ten zaś, który nie działa ze względu na cel, dla którego wyewoluował lub np. został zaprojektowany, jeżeli analizuje się artefakt, działa niepoprawnie. W tym sensie organizmy, które dążą do przetrwania, czyli unikają wzrostu entropii, można wyjaśnić, używając języka normatywnego. Zatem, w ujęciu Fristona, mechanizmy homeostatyczne są mechanizmami normatywnymi. Ta dyskusja prowadzi do, wydaje się, oczywistej konkluzji, która głosi, że dążenie do homeostazy nie jest jakimś pojedynczym mechanizmem albo funkcją organizmu, ale raczej zbiorem mechanizmów, które realizują określone funkcje. We fristonowskim ujęciu homeostazy mechanizm minimalizacji błędów predykcyjnych jest jednym z najbardziej podstawowych, a jego funkcja polega na generowaniu predykcji (por. Friston, Kiebel, 2009). Mechanizm ten jest realizowany przez model generatywny, który w standardowej interpretacji rozumiany jest jako mózg. Mechanizm ten, w ujęciu Fristona, jest jednym z zastosowań FEP. Nie będę jednak dalej rozwijał tego wątku, bowiem FEP, jak dowodziłem, nie ma zastosowania na gruncie PP. Godfrey-Smith zwraca jednak uwagę na istotne ograniczenie związane z mówieniem o mechanizmach homeostatycznych. Mianowicie: faktyczne mechanizmy homeostatyczne polegają na utrzymaniu stabilizacji tych organicznych własności, które nie są trywialnie związane z przetrwaniem. Znaczy to, że gdy jakaś złożona aktywność organizmu przyczynia się do przetrwania, jest ona prawdziwie homeostatyczna tylko wtedy, gdy istnieje jakaś pośrednia własność organiczna, czyli taka, że ta złożona aktywność przyczynia się do utrzymania homeostazy w tej pośredniej własności, o ile wnosi ona rzeczywisty wkład w przetrwanie. Przykładowo: inteligentne wykorzystanie ognia przez ludzi celem utrzymania ciepła ciała nie jest więc prawdziwym przypadkiem homeostazy realizowanej przez jakiś mechanizm poznawczy. Jest to zatem złożona zdolność organiczna, która umożliwia człowiekowi utrzymanie stałej temperatury ciała w zmiennych warunkach środowiskowych, umożliwiając tym samym przetrwanie w tych warunkach. Nie ma jednak powodu, by sądzić, że wszystko, co dzieje się za pośrednictwem poznania, jest w sposób trywialny homeostatyczne. Jeśli zatem skuteczna percepcja i koordynacja działań umożliwia organizmowi np. uniknięcie drapieżnika, to faktycznie nie jest to mechanizm

homeostatyczny. Pomaga ona co prawda organizmowi zachować podstawową integralność systemu organicznego. Nie byłoby to jednak możliwe, gdyby nie stabilność niektórych pośrednich własności organicznych, takich jak temperatura czy poziom cukru we krwi. Mechanizmy poznawcze są adaptacyjne, ale wyjaśnienie, dlaczego takie są, wymaga odwołania się do organicznych mechanizmów homeostatycznych (Godfrey-Smith, 1996, s. 79). Uwaga Godfreya-Smitha jest kluczowa dla naszej analizy mechanizmów homeostatycznych u Fristona. Dowodzi ona bowiem, że należy zachować dużą ostrożność podczas orzekania, że coś służy utrzymaniu homeostazy lub jest mechanizmem homeostatycznym.

W tym kontekście trzeba stwierdzić, że minimalizacji błędów predykcyjnych nie można uznać za faktyczny mechanizm homeostatyczny. Mechanizm ten może jedynie służyć przetrwaniu i działaniu danego organizmu w środowisku, o ile odwołuje się do organicznych mechanizmów homeostatycznych. W tym sensie minimalizacja błędów predykcyjnych jest sposobem realizacji homeostazy. Seth i Friston powiadają: „procesy poznawcze opierają się na podstawowych imperatywach ewolucyjnych w celu utrzymania fizjologicznej homeostazy" (Seth, Friston, 2016, s. 2). Jednak wbrew temu twierdzeniu samo wyjaśnienie mechanizmów homeostatycznych nie wystarcza, aby wyjaśnić, czym są mechanizmy minimalizacji błędów predykcyjnych[4]. Uwaga pokazuje ograniczony zasięg możliwości eksplanacyjnych modelu proponowanego przez Fristona. Niemniej pozwala nam ona zrozumieć to, że jeżeli chce się wyjaśnić normatywność mechanizmów predykcyjnych, to należy powiązać ją z normatywnością homeostatyczną.

Mówiąc o jakimś mechanizmie jako normatywnym, będę obecnie miał na myśli to, że realizuje on odpowiednią funkcję normatywną właściwą dla danego narządu, układu czy organizmu albo systemu poznawczego. Kiedy ją realizuje, wtedy działa poprawnie. Kiedy jej nie realizuje, wtedy działa niepoprawnie. W takim ujęciu określony mechanizm jest normatywny, ponieważ normatywna jest funkcja bądź funkcje, które dany mechanizm realizuje (por. Garson, 2013; por. także podrozdział *Normatywność a poziomy opisu*). Normatywność mechanizmu należy także

[4] Można jednak bronić podejścia Fristona, twierdząc, że tylko niektóre mechanizmy są homeostatyczne, zaś inne są allostatyczne. Allostaza, czyli „stabilność poprzez zmianę" (*stability through change*), jest procesem osiągania stabilności lub homeostazy na drodze zmian fizjologicznych i behawioralnych. Jest ona niezbędna dla utrzymania wewnętrznej żywotności organizmu w zmieniających się warunkach (por. Sterling, Eyer, 1988). Może zapewnić rekompensatę w reakcji na różne problemy organizmu (np. kompensuje w problemach z sercem, nerkami bądź wątrobą). Niestety, stany allostazy są nietrwałe i kompensacja może się szybko skończyć. Niektórzy zwolennicy FEP twierdzą, że skoro homeostaza nie pozwala na wyjaśnienie bogatej różnorodności procesów regulacyjnych, to należy odwołać się do allostazy. Ta ostatnia umożliwia proaktywne przygotowanie organizmu na potencjalne zmiany regulacyjne, przyczyniając się tym samym do minimalizowania energii swobodnej, *resp.* błędów predykcyjnych (por. Corcoran, Hohwy, 2017; Corcoran, Pezzullo, Hohwy, 2019).

Proponowane rozwiązanie z całą pewnością pozwala na lepsze i pełniejsze zastosowanie FEP do wyjaśniania wielu zjawisk. Nasuwa się jednak od razu ta sama obiekcja, którą sformułowałem w odniesieniu do mechanizmów homeostatycznych: należy zachować dużą ostrożność odnośnie do stwierdzenia, że jakiś mechanizm jest allostatyczny. Zapewne część mechanizmów introcepcyjnych można uznać za allostatyczne, nie oznacza to jednak, że model generatywny jako taki służy allostazie, czyli że da się satysfakcjonująco wyjaśnić za pomocą FEP.

uzasadnić poprzez odniesienie jej do podstawowego biologiczno-informacyjnego poziomu homeostatycznego, który wiąże się z podstawową tendencją organizmów żywych do przetrwania, tj. unikania okoliczności zwiększających entropię systemu. Innymi słowy, mechanizm jest normatywny, ponieważ odnosi się do takiego zachowania danego organizmu czy systemu poznawczego, jakie jest optymalne w określonym środowisku.

Pełne wyjaśnienie tej kwestii wymaga jednak odniesienia się do dyskusji nad tzw. funkcjami normatywnymi. Samo stwierdzenie, że mechanizm jest normatywny, ponieważ jest homeostatyczny, nie tylko jest trywialne, ale także może prowadzić do zanegowania normatywności mechanizmów, ponieważ dopóki nie wykaże się, że taka charakterystyka mechanizmów spełnia jakąś istotną funkcję eksplanacyjną, dopóty mówienie o ich normatywności należy uznać za przedwczesne, podobnie jak za przedwczesne uznałem mówienie o normatywności FEP[5]. Nie jest dla mnie w pełni satysfakcjonujące mówienie o normatywności mechanizmu albo funkcji tylko dlatego, że poprawnie lub niepoprawnie mogą realizować lub spełniać swoje zadania. Nietrywialne wyjaśnienie normatywności mechanizmów i funkcji musi wykazywać coś więcej. Musimy zatem przejść od tego, co w rozdziale *Przetwarzanie predykcyjne jako bayesowski model eksplanacyjny* określiłem jako preskryptywność epistemiczną, do preskryptywności ontycznej (por. podrozdział *Preskryptywizm epistemiczny i ontyczny*) w odniesieniu do mechanizmów normatywnych.

4.3. Pojęcie funkcji

Pojęcie funkcji jest wieloznaczne. Używa się go w bardzo wielu kontekstach, teoriach naukowych i filozoficznych, a także w języku potocznym. Stanowi ono integralną część sposobu myślenia zarówno w biologii, psychologii, jak i technice czy socjologii. Różne narządy, urządzenia, stanowiska, procesy mają przypisane określone funkcje, które mogą spełniać lub realizować. Niektórzy twierdzą, że pojęcie to jest niejasne z definicji, ponieważ wydaje się sugerować odniesienie do jakichś celów lub norm w przypadkach, gdy np. intencjonalność jest nieobecna; lub do skutków, gdy ich nie ma (np. gdy ma się do czynienia z dysfunkcją) (por. Krohs, Kroes, 2009, s. IX). Wydaje się jednak, że przyczyną tego stanu rzeczy nie jest sam pomysł wykorzystania pojęcia funkcji w wyjaśnieniach formułowanych na gruncie różnych nauk, ile pewien zamęt pojęciowy i właśnie brak jasnej eksplikacji tego, co się przez funkcje rozumie i w jakim celu stosuje się to pojęcie.

[5] Wielu autorów, pisząc o homeostazie albo mechanizmach biologicznych lub neuronalnych, zupełnie dobrze sobie radzi, nie korzystając ze słownika normatywnego. Celem niniejszych analiz jest wykazanie, że ten słownik również ma swoją moc eksplanacyjną.

W niniejszych rozważaniach będę odwoływał się do zastosowania tego pojęcia w filozofii biologii i kognitywistyce. Jednak również tutaj jego znaczenie nie jest jedno i istnieje wiele różnych jego wykładni. Eksplikacja znaczenia, którym się będę w moich analizach posługiwał, wymaga przedyskutowania propozycji kilku autorów. Na początek przyjmę dwie kluczowe przesłanki, które powinno spełniać interesujące mnie pojęcie funkcji: (1) pojęcie funkcji musi się w „jakimś stopniu" wiązać z pojęciem wyjaśniania – wymóg eksplanacyjny. Przez wymóg eksplanacyjny rozumiem, (i) że przypisanie danej funkcji określonemu systemowi pozwala na wyjaśnienie określonych efektów aktywności tego systemu (warunek słaby); oraz (ii) że przypisanie funkcji daje możliwość przyczynowego wyjaśnienia zjawiska, które jest realizowane za sprawą mechanizmu uzbrojonego w daną funkcję (warunek mocny); (2) pojęcie funkcji ma być tak skonstruowane, aby przy jego pomocy można było wyjaśnić dysfunkcję danego narządu, przedmiotu czy mechanizmu – wymóg normatywny. Takie pojęcie funkcji, jakie będzie spełniać obydwa te wymogi, pozwoli, jak sądzę, na wyjaśnienie natury mechanizmów predykcyjnych oraz ich normatywności[6].

Moje analizy w niniejszym podrozdziale rozpocznę od omówienia dwóch klasycznych teorii funkcji zaproponowanych przez Larry'ego Wrighta i Roberta Cumminsa. Następnie odwołam się do ujęcia Arno Woutersa, koncepcji funkcji właściwych Ruth Millikan i funkcji normatywnych Marka Bickharda. Na tle dyskusji z tymi koncepcjami sformułuję własną teorię funkcji, która zaspokoi wymogi eksplanacyjny i normatywny.

4.3.1. Funkcje w ujęciu Wrighta i Cumminsa

Pomyślmy o sercu zdrowego człowieka. Znajduje się ono w klatce piersiowej. Składa się z dwóch przedsionków i dwóch komór oraz zastawek. Ma określony cykl pracy, zwany „hemodynamicznym". Polega on na ciągłym przechodzeniu od stanów rozkurczu przedsionków i komór (tzw. pauza), przez skurcz przedsionków, skurcz komór (możliwy po zamknięciu tzw. zastawek), aż po ponowny rozkurcz. W efekcie tego cyklu serce pompuje krew. Jednocześnie zamknięcie zastawek wywołuje efekt akustyczny w postaci pierwszego tonu serca. Serce zatem pompuje krew i wydaje szereg prawidłowych zjawisk słuchowych (pierwszy i drugi ton serca, szmery przygodne itd.), a także nieprawidłowych (szmery wewnątrzsercowe, szmery pozasercowe itd.). W jaki sposób zatem należałoby wyjaśnić pracę serca? Czy poprzez odniesienie do funkcji pompowania krwi, czy może poprzez odwołanie do funkcji wydawania odpowiednich efektów akustycznych, a może poprzez od-

[6] Zaproponowane przeze mnie rozumienie mechanizmów oraz funkcji normatywnych ma spełniać zarówno warunek eksplanacyjny w sensie mocnym, jak i warunek normatywny.

niesienie do obu tych funkcji? Ktoś mógłby powiedzieć, że zdrowe serce pompuje krew. Byłoby to relatywnie dobre wyjaśnienie. Ale czy gorszym wyjaśnieniem byłoby stwierdzenie, że zdrowe serce wywołuje wiele określonych efektów akustycznych? Teoretycznie nie. Zatem które z tych wyjaśnień mówi nam coś o pracy zdrowego serca? Jedno z nich czy oba? A może żadne?

Przed tym problemem stanął Larry Wright. Jego odpowiedź była bardzo prosta. Zaproponował, aby funkcje danego narządu były odróżnialne od zwykłych efektów dzięki ich mocy eksplanacyjnej. Funkcja czegoś jest efektem, który wyjaśnia, dlaczego coś znajduje się tam, a nie gdzie indziej. I przykładowo serce znajduje się w klatce piersiowej, ponieważ w tym miejscu może pompować krew, a nie dlatego, że tutaj może wydawać pewne dźwięki. Zatem, stwierdza Wright, „funkcją *X-a* jest ta szczególna konsekwencja jej istnienia, która wyjaśnia, dlaczego *X* jest tam, gdzie jest" (1973). Funkcja *X-a* jest *Z* wtedy, gdy:

- *Z* jest efektem istnienia *X-a*;
- *X* jest tam, gdzie jest, ponieważ przyczynia się do tego *Z*.

Wright sądził, że jego koncepcja znajdzie zastosowanie zarówno w wyjaśnianiu funkcji artefaktów, jak i żywych organizmów. Są jednak sytuacje, w których tak rozumiana funkcja wyjaśnia nie te rzeczy, które byśmy chcieli, aby wyjaśniała. Przykładowo Christopher Boorse (1976) wykazał, że kiedy naukowiec widzi wyciek w wężu gazowym, to zanim będzie w stanie wyciek zatrzymać, może stracić przytomność od zatrucia. Według schematu Wrighta powinniśmy w takiej sytuacji powiedzieć, że funkcją przerwy w wężu jest uwalnianie gazu. Przerwa istnieje, ponieważ uwalnia gaz, unieruchamia naukowca, a wyciekający gaz jest konsekwencją przerwania węża. Godfrey-Smith (1993) stwierdza, że podejście Wrighta opiera się na pragmatyzmie w odniesieniu do wyjaśnień. Wright, inaczej niż np. Hempel (1965), podkreśla, że to, co określi się jako funkcję, zależy bezpośrednio od wyjaśnienia, a wyjaśnienie jest pragmatycznie wrażliwe na wiele sposobów. Ludzie wyjaśniają obecność rzeczy w kategoriach tego, co one robią. Funkcją jest zatem każdy efekt, który pomaga w takim wyjaśnieniu.

Funkcje Wrighta wiąże się z tzw. wczesną koncepcją etiologiczną[7]. Istotny jest w niej drugi warunek nałożony na funkcję, czyli stwierdzenie, że dana funkcja występuje, ponieważ przyczynia się do realizacji takiego, a nie innego działania. Jest to warunek powiązany z mechanizmem selekcji. Kontrowersje budzi warunek pierwszy, bowiem wymaga on, aby danemu organizmowi czy artefaktowi przypisać działania, które są efektem posiadania danej funkcji. Oznacza to, że np. dana cecha nie może mieć określonej funkcji, jeżeli określony organizm bądź artefakt nie powodowałby danego efektu. Przykładowo: funkcją serca jest pompowanie krwi tylko wtedy, gdy ono faktycznie tę czynność wykonuje. Czyli serce, które

[7] Wouters podkreśla jednak, że jest to bardzo słabe ujęcie etiologiczne, ponieważ nie wymaga ono, aby dana cecha została wybrana właśnie ze względu na jakąś określoną funkcję (Wouters, 2003, s. 634).

z powodu wypadku czy zawału obecnie nie może pompować samodzielnie krwi, w tym ujęciu takiej funkcji jest pozbawione. Oznacza to, że w praktyce nie jesteśmy w stanie odróżnić funkcji od dysfunkcji (Bielecka, 2018, s. 153).

Powraca w tym miejscu (analizowany już w podrozdziale *Przetwarzanie predykcyjne jako bayesizm oświecony*) problem normatywnych zobowiązań wyjaśnień. Wydaje się, że koncepcja Wrighta pozwala na wyjaśnienie funkcji, które nie korzysta ze słownika normatywnego, tylko odwołuje się do takich deskryptywnych pojęć jak efekt czy dystynkcja funkcja/wypadek, która pozwala mu na odróżnienie funkcji od korzyści losowej (por. Godfrey-Smith, 1993, s. 197). Niemniej w tym ujęciu dany organ, cecha czy artefakt określoną funkcję ma lub nie. Trudno jest zatem wyjaśnić, dlaczego coś może mieć funkcję, pomimo że obecnie nie realizuje przypisanej jej roli. Tak rozumianą funkcję należy z tego powodu raczej wiązać z systemowym ujęciem zaproponowanym przez Cumminsa (Bielecka, 2018, s. 152–153)[8] niż z ujęciami etiologicznymi podkreślającymi normatywną różnicę pomiędzy funkcją i dysfunkcją.

Cummins w artykule *Functional analysis* (1975) powiada, bezpośrednio atakując Wrighta, że błędem jest założenie, według którego odniesienie do skutków danej cechy biologicznej pozwala wyjaśnić jej obecność w danym organizmie. Zgadza się z tym, że program genetyczny wyjaśnia tego typu obecność, jednak nie decyduje o tym jakaś celowość procesu selekcji naturalnej, a przypadkowe mutacje. W wyjaśnianiu funkcji istotne jest zwrócenie uwagi na rolę, jaką dana funkcja gra w wyjaśnieniu określonej zdolności. Funkcje nie są efektami, które wyjaśniają, dlaczego coś istnieje, ale skutkami, które przyczyniają się do wyjaśnienia bardziej złożonych możliwości i dyspozycji systemu zawierającego dane zdolności czy cechy. Jeżeli zatem dany komponent systemu nie jest w stanie zrobić tego, co robią jego inne komponenty, które odgrywają określone role w wyjaśnianiu tego systemu, to komponent ten nie jest istotny i nie można go określić jako funkcję. Koncepcję Cumminsa nazywa się „systemową", ponieważ o tym, co nazwiemy funkcją, nie decyduje historia ewolucyjna bądź mechanizmy selekcji, ale rola, którą się jej przypisze w wyjaśnieniu danego systemu. Cummins (1975, s. 762[9]) powiada:

> *x* funkcjonuje jako φ w *s* (lub: *x* pełni funkcję φ w *s*) relatywnie do analitycznego wyjaśnienia *A* zdolności *s* do ψ, wtedy gdy *x* jest zdolne do φ-ania, w *s* oraz *A* poprawnie wyjaśnia zdolność *s* do wykonywania ψ między innymi przez odwołanie się do tego, że *x* jest zdolny do wykonywania φ w *s*.

Przykładowo: o tym, że serce ma funkcję narządu, który pompuje krew, decyduje to, że pompowanie przez niego krwi jest istotne dla funkcjonowania układu krążenia przenoszącego krew do innych narządów danego systemu. W tym

[8] Godfrey-Smith (1993) argumentuje, że – wbrew powszechnej opinii – koncepcji Wrighta i Cumminsa nie powinno się sobie przeciwstawiać, ale raczej wiązać je ze sobą z powodu przyjętych rozwiązań eksplanacyjnych.

[9] Cyt. za: Bielecka, 2018, s. 152.

ujęciu wyjaśnienie pracy serca w kategoriach jego funkcji do pompowania krwi jest poprawne, ponieważ organ wykonujący takie działania jest zdolny do tego, by dostarczyć krew do reszty organizmu. Funkcją zatem jest pewna dyspozycja, cecha czy zdolność określonego systemu, przy pomocy której można wyjaśnić jego przyczynową strukturę. Jest to ujęcie pragmatyczne, podobnie jak propozycja Wrighta. Zrywa ono jednak zupełnie z kwestiami dziedziczenia, ewolucyjnymi, oraz – co istotne – z wyjaśnieniami teleologicznymi. Wydaje się zatem, że jest to koncepcja nienormatywna.

Krytycy zarzucają jej niemożliwość wyjaśnienia dysfunkcji organizmu czy systemu oraz pragmatyzm ujęcia. Mianowicie: przypisanie funkcji danej zdolności jest zrelatywizowane do zainteresowań eksplanacyjnych badaczy (Millikan, 2011, s. 113). Systemy zatem nie muszą realizować określonych funkcji, jeżeli zaś im je przypiszemy, to tylko z powodu określonej strategii eksplanacyjnej, a nie dlatego, że dany system robi to, co robi, lub działa ze względu na określony cel. Millikan uważa, że problem przypisywania funkcji według Cumminsa ostatecznie jest problemem definicyjnym, ponieważ jej przypisanie zależy od wymogów, jakie nałoży się na daną dyspozycję. Innymi słowy, to w jaki sposób wyróżni się daną funkcję (a każda cecha biologiczna ma ich zazwyczaj kilka), jest uzależnione od poziomu szczegółowości przeprowadzanej analizy funkcjonalnej. Millikan podkreśla również, że Cummins w sposób niewłaściwy rozumie pojęcie selekcji. Program genetyczny jest istotny, ponieważ jak zaznacza Millikan (2011, s. 112–113):

> jakakolwiek cecha, która została wyselekcjonowana w celu spełnienia jej aktualnej funkcji, historycznie rzecz biorąc, *spowodowała* wzrost dostosowania. Właśnie o to chodzi, gdy mówi się, że cecha została wyselekcjonowana raczej *w jakimś celu* niż po prostu wyselekcjonowana. A to oznacza, że cecha ta musiała faktycznie w decydujący sposób rozstrzygać o przetrwaniu i/lub reprodukcji przynajmniej pewnych poprzedników posiadających tę cechę, i dlatego musiała pomóc wyjaśnić istnienie obecnego następcy posiadającego te cechę.

W praktyce oznacza to, że program genetyczny jest przyczyną pojawienia się danej funkcji. Nie oznacza to jednak, że jest on jej jedyną przyczyną (por. Neander, 1995).

Podsumujmy przeprowadzone analizy. Obie omówione koncepcje, tj. Wrighta i Cumminsa, nie spełniają wyróżnionego przeze mnie wymogu normatywnego, spełniając jednocześnie wymóg eksplanacyjny. Spełniają go jednak w ograniczony sposób. Mianowicie zakładają, że przypisanie funkcji jest zrelatywizowane do wyjaśnienia, a nie do faktycznych dyspozycji czy cech systemu. Jest to szczególnie widoczne u Cumminsa, który zupełnie odcina się od jakichkolwiek wyjaśnień etiologicznych. Wright w pewnym stopniu odwołuje się do historii narządu, niemniej przypisanie danej funkcji jest zrelatywizowane wyłącznie do tego, co dany

narząd faktycznie robi. Takie ujęcie nie wyjaśnia faktu posiadania przez dany system określonych funkcji, które z różnych powodów w danym momencie nie są realizowane. Uogólniając, należy stwierdzić, że proponowane tutaj podejście eksplanacyjne ma charakter epistemiczny, a nie ontyczny (por. Bechtel, 2011; Craver, 2012; Illari, 2013), co w praktyce oznacza, że wyjaśniane jest tylko to, co dany system robi, ale nie to, dlaczego to robi, czyli ma takie cechy, które przyczyniają się do określonych zdolności bądź działań (Millikan, 2011, s. 113). Z tego też powodu oba ujęcia, pomimo wielu zalet, jak i wpływu na szereg dyskusji w obrębie filozofii biologii i kognitywistyki, należy odrzucić jako potencjalne teorie wyjaśniające funkcje normatywne realizowane przez normatywne mechanizmy minimalizacji błędów predykcyjnych. Propozycja Millikan wydaje się dobrym kandydatem na taką teorię. Zanim jednak do niej przejdę, omówię wpierw propozycję wysuniętą przez Woutersa. Pozwoli nam ona uporządkować dyskusję nad funkcjami i ich normatywnością.

4.3.2. Funkcje w ujęciu Woutersa

Wouters (2003) wykazał, że należy wyróżnić cztery podstawowe sposoby, w jakich jest używane pojęcie funkcji w odniesieniu do badania żywych organizmów. Są to:

1. Funkcja jako aktywność (funkcja$_1$) – organizm jako taki, jego część, narząd, który coś robi lub jest w stanie robić.
2. Funkcja jako rola biologiczna (funkcja$_2$) – sposób, w jaki dany przedmiot lub działanie przyczynia się do jakiegoś złożonego działania lub zdolności organizmu.
3. Funkcja jako korzyść bądź przewaga biologiczna (*advantage*) (funkcja$_3$) – korzyść bądź przewaga (nad innymi organizmami) dla organizmu powiązana z obecnością pewnej własności bądź zachowania.
4. Funkcja jako wyselekcjonowany efekt (funkcja$_4$) – efekt, dla którego w przeszłości pewna cecha została wybrana i dzięki któremu można wyjaśnić jej obecność w populacji.

Przykład serca pozwoli lepiej zrozumieć różnicę pomiędzy poszczególnymi rodzajami funkcji. Skurcz serca jest czynnością, którą serce samo wykonuje (funkcja$_1$). Ta czynność umożliwia sercu spełnianie jego biologicznej roli (funkcja$_2$) w układzie krążenia, tj. pompowania krwi. Obecność układu krążenia, jako swoistego aktywnego środka transportu, ma tę przewagę biologiczną (funkcja$_3$) nad transportem pasywnym (np. dyfuzja), że tlen oraz inne gazy są transportowane w ten sposób znacznie szybciej. Dyfuzja nie jest wystarczająco szybka, aby dostarczyć tlen do wewnętrznych komórek organizmu w wymaganym tempie. Główną biologiczną zaletą (funkcja$_3$) posiadania jednego wyspecjalizowanego organu przeznaczonego do pompowania krwi (serce), zamiast bicia wszystkich poszczególnych naczyń

krwionośnych, jest wzrost wydajności systemu: system pompowania krwi przez wszystkie naczynia krwionośne nie byłby w stanie wytworzyć takiego ciśnienia, jak to, które jest generowane przez serce, oraz takiej samej mocy regulacyjnej, jak ta, którą serce ma. Pierwotna ewolucja serca kręgowców (funkcja$_4$) jest trudna do scharakteryzowania. Wiadomo jednak, że struktura serca kręgowców ulegała ogromnym zmianom w czasie ewolucji. Głównym bodźcem tych zmian było przyjęcie oddechu z powietrza przez kręgowce podobne do ryb około 350 milionów lat temu, a następnie pojawienie się zwiększających się wymagań dotyczących wydajności systemu transportowego, powiązanych z postępującym rozwojem różnych kręgowców (Wouters, 2003, s. 635).

Funkcja$_1$ to wszelka aktywność danego organizmu. Funkcja$_2$ wiąże się z tym, w jaki sposób dana cecha lub zachowanie jest wykorzystywane przez organizm lub jego części. Rola biologiczna określa, jaki wkład wnosi określona cecha bądź aktywność do powstania złożonych własności organizmu, tj. takich własności, jakie nie są po prostu sumą części ich własności, ale wynikiem sposobu, w jaki te części i ich działania są powiązane (Wouters, 2003, s. 638; por. także Craver, 2001). Przykładem takich własności jest dążenie do przeżycia, zachowania homeostazy. Biolog, formułując wyjaśnienie funkcjonalne, dzieli organizm na określone podsystemy (np. układ nerwowy, układ krążenia), z których każdy jest istotny dla pojawienia się określonych zachowań oraz przeżycia organizmu (Juś, 2014, s. 210). Zdania przypisujące określone role biologiczne mają następującą postać: „Przedmiot lub zachowanie *i* pełni biologiczną funkcję *f* w wytwarzaniu aktywności lub zdolności *c* systemu *s* w życiu organizmów *o*" (Wouters, 2003, s. 639). Przypisywanie roli biologicznej jest istotne z perspektywy wyjaśnień funkcjonalnych bądź mechanistycznych (por. Bechtel, 2007; Craver, 2001, 2008; por. także Cummins, 1975), co wiąże się z faktem, że można takie stwierdzenia formułować na różnych poziomach organizacji (od chemicznych, przez fizjologiczne, aż po behawioralne). Należy dodać, że role biologiczne są z definicji takimi działaniami organizmów, jakie opisywalne są przy użyciu predykatów odnoszących się do środowiska danego organizmu. Te, które nie odnoszą się do środowiska, oznaczają określone funkcje$_1$. Oznacza to, że dana funkcja$_1$ może odgrywać różne role biologiczne (funkcja$_2$), w zależności od gatunku (por. Bock, Wahlert, 1965). Różnica pomiędzy funkcją$_1$ a funkcją$_2$ jest różnicą pomiędzy własnościami, które dany przedmiot ma, a sposobem, w jaki dany przedmiot pomaga wytworzyć określoną umiejętność systemu, którego stanowi część. W przypadku serca jego rytmiczne bicie jest funkcją$_1$, czyli aktywnością. Pompowanie krwi zaś jest rolą biologiczną serca (funkcją$_2$), czyli wkładem, jaki serce wnosi w pracę układu krążenia. Role biologiczne możemy przypisywać określonym indywiduom, tj. indywidualnym zachowaniom bądź cechom obecnym w danym organizmie; lub możemy je przypisywać określonej klasie organizmów na drodze empirycznej generalizacji (Wouters, 2003, s. 640–642).

Funkcja$_3$, czyli przewaga biologiczna, wiąże się z taką zdolnością bądź cechą organizmu, dzięki jakiej ma on większe szanse na przeżycie niż organizm, który danej zdolności bądź cechy nie ma. Wouters, analizując literaturę przedmiotu, podkreśla, że twierdzenia dotyczące przewagi biologicznej (funkcja$_3$) poprzedzone są często przypisaniem określonym własnościom jakichś ról biologicznych (funkcja$_2$). Z tego to powodu proponuje, aby posiadanie roli biologicznej przypisywać raczej przedmiotom i zachowaniom, zaś z cechami należy wiązać źródło przewagi biologicznej.

Funkcja rozumiana jako wyselekcjonowany efekt (funkcja$_4$) wiąże się z przypisaniem takiej cechy, jaka w przeszłości wykształciła się na drodze ewolucyjnych mechanizmów doboru naturalnego. Historyczne ujęcie funkcji opiera się zatem na stwierdzeniu, że funkcje danej cechy bądź zachowania nie są determinowane przez aktualne uwarunkowania, ale przemiany ewolucyjne oraz przeszłe oddziaływania środowiskowe.

Czy i które z wyróżnionych przez Woutersa rodzajów funkcji biologicznych spełniają postawione przeze mnie wymogi eksplanacyjny i normatywny? Jeżeli chodzi o funkcję$_1$, czyli tę, która odnosi się do poszczególnych aktywności organizmu bądź jego wybranych części, to trudno oczekiwać, żeby przypisanie predykatu „bycia funkcją" dla dowolnej aktywności systemu dawało jakiś duży zysk poznawczy. Wouters raczej relacjonuje tutaj jeden ze sposobów, w jaki badacze używają tego pojęcia, aniżeli przypisuje mu jakieś większe znaczenie eksplanacyjne. W takim ujęciu funkcjami serca są zarówno rytmiczne pompowanie krwi, jak i generowanie efektów akustycznych. W związku z tym uważam, że funkcja$_1$ nie spełnia żadnego z postulowanych przeze mnie wymogów. W takim razie zastosowanie jej do wyjaśnienia natury mechanizmów predykcyjnych wydaje się trywialne, ponieważ mówi ono tylko tyle, że mechanizmy te minimalizują błędy predykcyjne. Inaczej sprawa wygląda z funkcjami$_2$ i $_3$, czyli rolą i przewagą biologiczną. Twierdzę, że należy powiązać ze sobą te funkcje, ponieważ jedna i druga odnoszą się do tego samego przedmiotu, niemniej z dwóch różnych stron. W odniesieniu do interesujących nas mechanizmów minimalizacji błędów predykcyjnych zastosowanie obu tych funkcji prezentuje się następująco: zgodnie z przyjętą w tej pracy wykładnią, niektóre mechanizmy minimalizacji błędów predykcyjnych służą przetrwaniu i działaniu danego organizmu w środowisku, o ile są one sposobem realizacji biologicznych mechanizmów homeostatycznych[10]. Zatem funkcją$_2$ (rolą biologiczną) mechanizmów predykcyjnych jest dążenie do utrzymania homeostazy, czyli minimalizacja niepewności. Innymi słowy, dany mechanizm predykcyjny odgrywa biologiczną rolę w *układzie nerwowym*, polegającą na wytwarzaniu zdolności *do utrzymywania homeostazy* systemu, w życiu takich, a nie innych organizmów (tj. tych, które implementują bayesowski model generatywny). Funkcja$_3$ (przewaga

[10] Oznacza to, że niektóre mechanizmy predykcyjne są powiązane z wymogami homeostatycznymi, inne zaś nie lub tylko pośrednio.

biologiczna) mechanizmów predykcyjnych polega na tym, że organizmy, które są *uzbrojone* w model generatywny realizujący takie mechanizmy, mają większe szanse na przeżycie niż organizmy bądź systemy, które takich mechanizmów nie realizują w ogóle bądź ich nie realizują z jakiegoś powodu (np. uszkodzenie układu nerwowego, choroba psychiczna). Funkcja$_4$ jest o tyle istotna, o ile dzięki jej przypisaniu możemy powiedzieć, że mechanizmy minimalizacji błędów predykcyjnych są jednym ze sposobów, w jaki niektóre organizmy wyewoluowały i zaadaptowały się w środowisku (por. Hohwy, 2015a, 2015b; Pezzulo, 2017; por. także Friston, Stephan, 2007). Pełne znaczenie tego rodzaju funkcji omówię przy okazji analizy pojęcia funkcji właściwej Millikan. W tym momencie należy powiedzieć, że funkcja$_4$ wyjaśnia nam sposób, w jaki funkcje$_{1, 2, 3}$ przyjęły taką, a nie inną formę.

Wiadomo już, że przypisywanie funkcji$_1$ jest opisowo istotne, niemniej eksplanacyjnie niewiele nam daje. Inaczej jest w przypadku pozostałych funkcji. Ich wykorzystanie jest kluczowe do sformułowania nietrywialnych wyjaśnień funkcjonalnych, mechanistycznych oraz istotnych dla nauk biologicznych wyjaśnień etiologicznych (Juś, 2011, s. 221; Wouters, 2003, s. 665–658)[11]. Jak sprawa wygląda z wymogiem normatywnym? Trudno jednoznacznie stwierdzić, ponieważ możliwość dysfunkcji bądź nieprawidłowo realizowanej funkcji nie jest problemem, który Woutersa zajmuje. Jak wspomniałem, tworzy on raczej katalog możliwych użyć tego pojęcia, nie wchodząc w różne niuanse. Niemniej wydaje się, że w przypadku funkcji$_{2, 3, 4}$ można ująć je od strony normatywnej. Stanie się to jaśniejsze przy okazji omawiania propozycji Millikan i Bickharda. Jeżeli chodzi o funkcję rozumianą jako aktywność (funkcja$_1$), to albo daną funkcję się przypisuje, albo nie. Nie ma tutaj miejsca na możliwość dysfunkcji. W tym sensie mówienie o funkcji jako aktywności zbliża nas do systemowego ujęcia Cumminsa.

Podsumowując, dyskusja nad czterema rodzajami funkcji wyróżnionymi przez Woutersa pozwala na rozpoznanie istotnych elementów, które nietrywialna eksplanacyjnie teoria spełniająca wymóg normatywny musi zawierać:

- odwołanie do roli biologicznej, którą dana funkcja spełnia;
- odwołanie do przewagi biologicznej danej funkcji, która wynika ze spełnianej przez nią roli biologicznej;
- odwołanie do etiologii danej funkcji, która wyjaśnia genezę pojawienia się odpowiedniej roli i przewagi biologicznej;
- możliwość wyjaśnienia dysfunkcji, czyli tego, dlaczego dana funkcja nie jest realizowana albo jest realizowana nieprawidłowo, pomimo że dany system ją ma i teoretycznie może ją realizować.

[11] Na pytanie o to, czy te pojęcia spełniają warunek eksplanacyjny w sensie słabym czy mocnym, odpowiem w podrozdziale *Funkcje w ujęciu Millikan* przy okazji omawiania koncepcji Millikan.

4.3.3. Funkcje w ujęciu Millikan

Millikan w ramach swojej koncepcji z zakresu teleosemantyki rozwija etiologiczną teorię funkcji właściwej (1984, 1989, 2011). Pojęcie to wiąże się z reprodukowanymi lub powielanymi rzeczami bądź artefaktami, których poprzedniki pomagają wyjaśnić ich przetrwanie na drodze ciągłej reprodukcji. Innymi słowy, przedmiot ma funkcję właściwą, jeżeli wywodzi się z linii, która zawdzięcza swoje przetrwanie istnieniu korelacji pomiędzy odróżniającymi ją cechami a efektami, które można określić jako funkcje tych cech[12] (Millikan, 1989, s. 288–289, 2011, s. 111). Oznacza to, że funkcja właściwa przysługuje takim cechom, jakie zostały wyselekcjonowane na drodze mechanizmów ewolucyjnego doboru naturalnego. Wiąże się ona zatem bezpośrednio z funkcją$_4$ Woutersa. Millikan jednak nakłada dodatkowy warunek, który czyni jej propozycję znacznie subtelniejszą i eksplanacyjnie nietrywialną. Mianowicie stwierdza, że cecha, która ma funkcję właściwą, nie została wyselekcjonowana tak po prostu, ale *w pewnym celu* (por. także Sober, 1984). Z tego też powodu zamiast określać funkcje jako właściwe, można je również nazywać „celowymi" (Millikan, 1989, s. 291). Dlaczego ta modyfikacja ujęcia Woutersa jest taka istotna? Jest istotna, ponieważ pozwala przypisywać funkcje tym systemom, organizmom czy artefaktom, które pomimo że mają swoje funkcje, to z jakiegoś powodu nie są zdolne do ich realizowania. Owe przedmioty mogą nie realizować swoich funkcji właściwych z powodu pewnego „uszkodzenia" lub ze względu na określone warunki tła, np. te, które pomogły utrzymać działanie danej funkcji u poprzedników, a teraz są nieobecne. W takiej sytuacji mogą one nie działać właściwie, pomimo że nie pojawia się tutaj żadna anormalność. Funkcje właściwe są sterowane przez takie systemy percepcyjne i (lub) poznawcze, których zadaniem jest optymalna zmiana relacji organizm – środowisko tak, aby środowisko dostarczało najkorzystniejszych warunków otoczenia. Niektóre z funkcji dotyczą zmiany w obrębie środowiska, w taki sposób, aby dopasowało się ono do organizmu; albo odwrotnie: część z nich dotyczy zmiany organizmu tak, żeby ten dopasował się do środowiska. Celem systemów percepcyjnych i poznawczych jest sprawienie, aby organizm i środowisko pozostały w określonej, optymalnej relacji do siebie. Można zatem mówić, że celem funkcjonowania systemów percepcyjnych i poznawczych jest wytworzenie pewnych struktur relacyjnych (Millikan, 2011, s. 122–123).

Funkcje Millikan mogą wygasać oraz ewoluować w czasie, tzn. zmieniać się w tzw. funkcje pochodne. Pojęcie funkcji pochodnych autorka objaśnia na przykładzie kameleona. Funkcją pochodną od funkcji właściwej dopasowywania koloru skóry przez kameleona do koloru tła jest funkcja zmieniania koloru skóry na zielony w momencie, gdy zwierzę usiądzie na zielonym liściu bądź gałęzi. Funkcja

[12] W korelacji przeciwstawia się przypadki pozytywne i negatywne.

właściwa to zmiana koloru skóry, zaś funkcja pochodna to zmiana koloru skóry na zielony w odpowiednim środowisku (Millikan, 1984, rozdział 2). Funkcje pochodne umożliwiają powielanie struktury relacyjnej funkcji właściwych. Można zatem powiedzieć, dookreślając wyjściowe ujęcie, że funkcje są wynikiem historii naturalnej, podczas której zostały wyselekcjonowane w reakcji na określoną cechę, która sprzyjała ich przetrwaniu i reprodukcji. Z tego powodu należy stwierdzić, że nośnikiem funkcji są określone typy (np. gatunki), a wtórnie indywidua, czyli egzemplarze danego typu (Millikan, 2011, s. 121–123; por. także Bielecka, 2018, s. 172–173).

Millikan swoje ujęcie funkcji wiążę po pierwsze z tym, jak dana rzecz została zaprojektowana lub jak działa celowo, w przeciwieństwie do tego, co robi przypadkowo; po drugie zaś z istnieniem pewnego rodzaju „wzorca", który „powtarza się w wielu formach, na wielu poziomach i w wielu dziedzinach, oraz który – jak się wydaje – można odnaleźć wszędzie tam, gdzie w naturalny sposób przypisuje się cel i/lub intencjonalność" (Millikan, 2011, s. 109). Z tego też powodu ujęcie to określa mianem „normatywnego", zastrzegając jednocześnie, że takie rozumienie normatywności nie ma nic wspólnego z jakimś „ocenianiem" czy „wartościowaniem". Pojęć normatywnych używa się do wskazywania takich miar, od jakich są możliwe rzeczywiste odstępstwa (Millikan, 2011, s. 110–111). W takim ujęciu pojęcie funkcji właściwej należy traktować jako swoistą miarę czy normę dla orzekania, czy coś jest funkcją, czy nią nie jest, lub – co istotne – czy być może ma się do czynienia z dysfunkcją danego systemu, organizmu lub artefaktu. Gdzie indziej Millikan dodaje, że funkcja właściwa jest to taka funkcja, co do której „zakłada się", że rzeczy ją realizują (Millikan, 1984).

4.3.3.1. Funkcje właściwe a mechanizmy predykcyjne

Powróćmy na chwilę do analiz przeprowadzonych przeze mnie w podrozdziale *Przetwarzanie predykcyjne jako bayesizm oświecony*. Zadałem tam pytanie, czy słownik normatywny jest istotny dla wyjaśnienia funkcjonowania mechanizmów, które realizują określone funkcje? Argumenty Millikan pozwalają odpowiedzieć na powyższe pytanie twierdząco. Autorka ta wykazuje nie tylko, że wyjaśnienie odpowiedniej funkcji zakłada użycie takich pojęć normatywnych jak dysfunkcja czy wyjaśnienie oraz warunki Normalne[13], ale również że funkcje *per se* są normatywne dla funkcjonowania danego organizmu czy organu lub którejś z jego części. Należy zatem powiedzieć, że jej ujęcie daje argumenty na rzecz stanowiska, które

[13] Warunki te dotyczą właściwego funkcjonowania danego organu oraz regularnych zależności z innymi częściami organizmu i środowiskiem. Millikan pisze o warunkach i wyjaśnieniach Normalnych dużą literą, aby podkreślić, że są to warunki i wyjaśnienia istotne, nie zaś typowe dla realizacji danej funkcji (Bielecka, 2018, s. 176–177).

określiłem jako „preskryptywizm epistemiczny" w odniesieniu do bayesowskiego PP. Stanowisko to głosi, że sama struktura i funkcjonalność bayesowskiego modelu generatywnego jest normatywna dla percepcji oraz działań podejmowanych przez dany organizm, ponieważ mechanizmy odpowiedzialne za realizację tych funkcji są bayesowskie. Oczywiście analizy Millikan dotyczą innej grupy zagadnień. Istnieje jednak możliwość, że gdybyśmy chcieli zastosować jej rozwiązanie w odniesieniu do PP, to można by wykazać, że minimalizacja błędów predykcyjnych jest funkcją właściwą modelu generatywnego. Mechanizm predykcyjny, implementując regułę Bayesa, minimalizuje potencjalne błędy predykcyjne. Wyjaśnienie jego funkcjonowania odwoływałoby się do wykazania jego roli i przewagi biologicznej oraz jego etiologii, która następnie wyjaśniałaby genezę pojawienia się odpowiedniej roli i przewagi biologicznej, a także wyjaśniałaby możliwość dysfunkcji tego mechanizmu. Jeżeli chodzi o wyjaśnienie etiologiczne, to można odwołać się do badań nad układem nerwowym i jego ewolucją. Wiadomo, że zrekonstruowanie ewolucji układu nerwowego z wielu powodów jest zadaniem bardzo trudnym (np. z powodu braku dostępu do bogatego zapisu kopalnego). Istnieją niemniej dwie główne hipotezy dotyczące pierwotnej funkcji układu nerwowego. Według pierwszej z nich podstawową funkcją tego układu było utrzymanie łączności pomiędzy percepcją a działaniem. Według drugiej funkcją tą było formowanie działań (por. Godfrey-Smith, 2018, s. 39–40; Keizer, van Duijn, Lyon, 2013). Obie te wizje w pewnym sensie pozwalają uzasadnić rolę biologiczną modelu generatywnego minimalizującego błędy predykcyjne, a także przewagę biologiczną organizmów uzbrojonych w ten model nad tymi, które tego modelu nie mają[14]. Najbardziej jednak sporną kwestią jest pytanie o to, kiedy i w jaki sposób mechanizm predykcyjny może stać się dysfunkcyjny. Innymi słowy, co decyduje o tym, że mechanizm ten można uznać za normatywny? Wstępna, zasugerowana przeze mnie w rozdziale *Przetwarzanie predykcyjne jako bayesowski model eksplanacyjny*, odpowiedź mówi, że mechanizm jest normatywny, wtedy gdy realizuje on odpowiednią funkcję normatywną dla danego narządu czy układu. W tym sensie mechanizm minimalizacji błędów predykcyjnych jest normatywny, jeżeli może błędnie minimalizować swoje błędy.

Co to znaczy? Wyobraźmy sobie mikser kuchenny. Taki mikser realizuje swoją funkcję np. wtedy, kiedy miksuje jajka z mąką. Mówiąc językiem Millikan, należy stwierdzić, że funkcją właściwą (normatywną) miksera jest miksowanie. Dlaczego jest to funkcja właściwa? Ponieważ stanowi ona miarę dla tego, co określimy jako nieprawidłowe działanie miksera albo jego uszkodzenie. Aby orzekać o danym urządzeniu, że „nie działa" albo „działa prawidłowo", należy wiedzieć, na czym polega jego funkcja właściwa, tj. funkcja, która (w tym przypadku) została dla

[14] Nie jest do końca jasne, czy model generatywny jest przynależny człowiekowi, czy też naczelnym, ssakom, czy w ogóle organizmom żywym. Badacze nie podają minimalnych warunków, jakie musi spełniać system uzbrojony w ten model.

danego urządzenia zaplanowana przez odpowiedniego projektanta. Zatem to brak możliwości realizacji danej funkcji jest warunkiem koniecznym dla „niedziałania" danego mechanizmu, a nie brak funkcji, jakby mogli stwierdzić zwolennicy teorii systemowych. Nie jest to jednak warunek wystarczający (np. niepodłączony do prądu mikser nie może wykonywać swojej funkcji, co nie oznacza od razu, że jest on uszkodzony). W proponowanym ujęciu dany mechanizm jest normatywny, ponieważ spełnia warunki, które muszą zajść, aby dane działanie (poznawcze lub pozapoznawcze) zostało zrealizowane. Stwierdzenie, że określony mechanizm jest normatywny, oznacza po prostu, że jest możliwe, aby dany system *X* był mechanizmem dla aktywności *Y*, pomimo faktu, że (np. w danym momencie) *X* nie może wykonać *Y* (por. Garson, 2013)[15]. W odniesieniu do PP należałoby powiedzieć: mechanizm minimalizacji błędów predykcyjnych jest normatywny, ponieważ jest możliwe, aby model generatywny był mechanizmem dla aktywności, jaką jest minimalizacja błędów predykcyjnych, pomimo faktu, że w danym momencie model ten nie może minimalizować tych błędów lub minimalizuje je niepoprawnie. Wcześniej stwierdziłem, że normatywność mechanizmu predykcyjnego należy uzasadnić poprzez odniesienie jej do podstawowego biologiczno-informacyjnego poziomu homeostatycznego, czyli do takiego zachowania danego organizmu czy systemu poznawczego, jakie jest optymalne w określonym środowisku. Wydaje się, że przyjęcie koncepcji funkcji właściwej Millikan satysfakcjonująco spełnia ten wymóg. Mianowicie: etiologiczne ujęcie funkcji po pierwsze tłumaczy jej rolę i przewagę biologiczną, po drugie zaś pokazuje (w jakimś stopniu) jej ewolucyjne pochodzenie.

Poważnym mankamentem ujęcia Millikan jest jednak to, że w pierwszym rzędzie przypisuje się tutaj funkcje typom biologicznym (np. gatunkowi człowieka), a wtórnie egzemplarzom (np. określonej osobie żyjącej teraz w takim, a nie innym środowisku) (Millikan, 2011, s. 121–123). Jest to pewnego rodzaju słabość, ponieważ w takim ujęciu normatywność jest przede wszystkim analizowana na poziomie gatunkowym, a nie indywidualnym. W praktyce oznacza to, że określenie ogólnej reguły dotyczącej np. dysfunkcji modelu generatywnego, nie pozwoli nam wyjaśnić, dlaczego np. dane indywiduum zachowało się w taki, a nie inny sposób. Mam na myśli to, że jeżeli PP ma być nie tylko formalnym i abstrakcyjnym sposobem wyjaśniania i mówienia o umyśle, percepcji i ich związkach z działaniem (por. Clark, 2013b; Seth, 2015), to należy wykazać sposób, w jaki model PP pozwala wyjaśnić zakorzenione w świecie, nie tylko biologicznym, ale przede wszystkim materialnie i kulturowo ustrukturyzowanym, interakcje pomiędzy konkretnym, ucieleśnionym indywiduum a jego środowiskiem, które – dodajmy – nie jest pozbawione znaczenia, ale pełne możliwości działań, wartości, potencjalnych zagrożeń

[15] Należy jednak podkreślić, że Garson nie mówi o mechanizmach normatywnych, ale funkcjonalnych, które realizują funkcje normatywne.

i nagród etc. Chodzi o znalezienie zatem takiego ujęcia normatywności, które sprosta temu wyzwaniu. Wydaje się, że propozycja Millikan ze względu na swój historyczny charakter takiego ujęcia normatywności nie oferuje.

4.3.3.2. Krytyka podejścia teleosemantycznego

Podejście Millikan jest również przedmiotem ataku tych autorów, którzy twierdzą, że nie potrzeba powoływać się na pojęcie normatywności, aby wyjaśnić funkcje. Przykładowo Paul Davies (2001) twierdzi, że chociaż większość cech naturalnych rzeczywiście ma określone właściwości funkcjonalne, to właściwości te nie są wytworzone i uprzywilejowane przez procesy selekcji naturalnej. Ta ostatnia, zdaniem Daviesa, nie jest istotna dla wywyższenia jakiejkolwiek funkcji. Aby wyjaśnić funkcję, należy zaproponować taką teorię, jaka – nie powołując się na pojęcia teleologiczne i normatywne – w bardziej satysfakcjonujący sposób pasuje do naturalistycznego poglądu na świat (*naturalistic worldview*) (Davies, 2001, s. XIII). Davies proponuje, aby przyjąć teorię systemową, która konceptualizuje funkcje jako efekty komponentów systemowych, przyczyniających się do zwiększenia ogólnych możliwości systemu, do którego te komponenty należą. W ujęciu tym kładzie się nacisk na strukturę i organizację systemów naturalnych oraz pracę wykonywaną przez ich komponenty. Przykładowo zadaniem serca ssaka jest pompowanie krwi, ponieważ pompowanie w znaczący sposób przyczynia się do zwiększenia wydajności układu krążenia. Funkcje są zatem istotne z punktu widzenia systemowego. Ich historia ewolucyjna nie jest istotna. Davies zgadza się, że każda funkcja ma swoją etiologię, niemniej możemy sobie wyobrazić, że pompowanie krwi wciąż byłoby funkcją serca, nawet gdyby jakieś inne urządzenie bądź organ zastąpiło je na początku ewolucji ssaków. Należy w takim razie stwierdzić, że funkcje są istotnym wkładem w określone zdolności systemu, podczas gdy selekcja naturalna może je jedynie zachować, zmodyfikować bądź wyeliminować. Z całą pewnością, konkluduje Davies, nie jest ona ich źródłem i nie można ich uzasadnić poprzez odniesienie do ich historii.

Jak zatem systemowo wyjaśnić dysfunkcję? Davies zgadza się z przekonaniem, które głosi, że niektóre organizmy lub ich cechy są lepiej przystosowane niż inne. Nie zgadza się jednak z konstatacją, że istnieją jakieś „normy natury", które zostały narzucone przez przodków w wyniku ich sukcesu adaptacyjnego. Przez normy natury rozumie on te, które utrzymują organizm w pewnej równowadze i które umożliwiają pojawienie się jakichś wadliwych działań czy zachowań. Założenie, że takie normy istnieją, sprawia, zdaniem Daviesa, że przypisuje się istotom żywym określone formy celowości lub bycia zaprojektowanymi, co znacznie wykracza poza wyjaśnienia, jakie oferuje rozumiana po darwinowsku teoria adaptacji. Jak wiemy,

systemowym ujęciom funkcji można zarzucić, że nie radzą sobie z wyjaśnieniem dysfunkcji. Autor *Norms of nature* zgadza się z takim zastrzeżeniem, przy czym uważa, to za pewną korzyść przemawiającą na rzecz podejścia systemowego, a nie za jego słabość. Argumentuje następująco: przypisanie nieprawidłowości w działaniu naturalnym, a nie inżynieryjnym, cechom czy organizmom jest po pierwsze sprzeczne z naturalistycznym podejściem do badań w ogóle[16], po drugie jest sprzeczne z teorią ewolucji przez dobór naturalny (Davies, 2001, s. 29). Zdaniem Daviesa coś takiego jak historyczne dysfunkcje jest niemożliwe, ponieważ z punktu widzenia procesów selekcji nie ma „lepszego" bądź „gorszego" funkcjonowania danego organu czy cechy. Nie ma bowiem czegoś takiego jak dysfunkcja, która da się uzasadnić poprzez odniesienie do historii selekcji danego organu[17]. Nie istnieją również dysfunkcje systemowe. Funkcje są definiowane w odniesieniu do systemowych zdolności, które z kolei przyczyniają się do innych systemowych zdolności. Zatem funkcje, które z jakiegoś powodu nie spełniają swoich ról w systemie, nie mogą być zaliczone do danego typu, ponieważ – na jakiej innej podstawie można je przypisać do danego typu, jeżeli nie spełniają swojej systemowej roli? Ze względów badawczych możemy traktować takie funkcje jako te, które nie wywiązują się ze swojej roli, jest to jednak zależne od naszej strategii badawczej lub podejścia, a nie ze względu na charakter samej funkcji (Davies, 2001, s. 212–213).

Na podstawie powyższych analiz Davies stwierdza, że nie ma czegoś takiego jak normatywność funkcji czy – szerzej ujmując – normatywność w przyrodzie. Odrzuca on więc pogląd, który określiłem jako „preskryptywizm ontyczny". Można mu za to przypisać bardzo słabą wersję preskryptywizmu epistemicznego, zgodnie z którą mówienie o funkcjach jako normatywnych jest przydatne ze względów czysto pragmatycznych. Stwierdzenie, że dana funkcja jest normatywna, może odzwierciedlać nasze oczekiwania, które wiązać należy z normami predykcyjnymi nauki oraz normami eksplanacyjnymi (por. Craver, Kaplan, 2018). Należy jednak przełamać „nawyk naszego" umysłu, który polega na tym, że konceptualizuje się żywy świat w kategoriach projektowania, w ten sposób, że przypisuje się jakieś normy do cech naturalnych (Davies, 2001, s. 214–215). Propozycja Daviesa stanowi radykalizację stanowiska empiryzmu w modelowaniu czynności umysłowych bronionego przez Elqayam i Evana (2011; por. podrozdział *Normatywność a poziomy opisu*). Po pierwsze jego podejście jest wolne od słabości tego stanowiska w tym sensie, że porzuca „naiwne" rozumienie normatywności proponowane przez Elqayam i Evansa, w ten sposób, że wykorzystuje nietrywialne ujęcie nor-

[16] Teoria naturalistyczna funkcji, która stanowi część naturalistycznego podejścia do badań, ma opierać się w ujęciu Daviesa na stwierdzeniu, że funkcje powinny być testowalne empirycznie, oraz na przekonaniu, że wiarygodna teoria funkcji powinna wyjaśniać kryteria stosowania tego pojęcia w biologii, niezależnie od tego, co biolodzy na ich temat myślą.

[17] Argumenty na rzecz tej tezy Davies formułuje w siódmym rozdziale swojej książki (por. Davies, 2001, s. 203–206).

matywności biologicznej; po drugie zaś pokazuje określony sposób podejścia, który faktycznie rozwiązuje pewne problemy badawcze bez powoływania się na pojęcia normatywne, jedynie wykorzystując ujęcie systemowe. Strategia Daviesa opiera się na wykazaniu, że zwolennicy ujęć teleosemantycznych i etiologicznych w odniesieniu do funkcji popełniają błąd, który polega na tym, że myli się tutaj perspektywę systemu z perspektywą badacza tego systemu. Z perspektywy systemu nie jest istotne, czy danej funkcji przypisze się normatywność. Normatywność funkcji jest istotna dla badacza, który chce daną funkcję wyjaśnić jako celową lub w jakiś sposób zaprojektowaną (por. także Kolchinsky, Wolpert, 2018, s. 2).

Przyjęcie krytyki Daviesa może sugerować, że ujęcie Millikan należy do preskryptywizmu epistemicznego, czyli że istotny dla mojego ujęcia wymóg normatywny spełnia jedynie w słabym sensie. Millikan bowiem odwołuje się do słownika normatywnego, uzasadniając to etiologią typu odpowiednich funkcji, ale ostatecznie nie wykazuje zależności normatywnych, czyli nie daje mocnego argumentu na rzecz tezy, że funkcje są normatywne również systemowo, a nie ze względu na ich historię. Można oczywiście nie zgodzić się z krytyką Daviesa i bronić historycznego ujęcia normatywności funkcji[18], niemniej nawet przyjmując argumenty Millikan jako zasadne, wciąż pozostaje problem, o którym wspomniałem wcześniej, dotyczący różnicy pomiędzy normatywnością rozważaną na poziomie gatunku a normatywnością w odniesieniu do indywiduum. Krytyka Daviesa pozwala również na stwierdzenie, że zarówno propozycja Woutersa, jak i Millikan spełnia wymóg eksplanacyjny w sensie słabym. Oznacza to, że przypisanie danej funkcji jakiemuś systemowi pozwala wyjaśnić takie, a nie inne efekty jego aktywności, ale nie pozwala na przyczynowe wyjaśnienie danego zjawiska, ponieważ przypisanie funkcji jest wyłącznie zrelatywizowane do zainteresowań badaczy bądź obranej przez nich strategii eksplanacyjnej.

Krytyka Daviesa wykazuje przekonująco, że wszelkie próby mówienia o normatywności funkcji są skazane na porażkę. Jest to konkluzja, która w pewnym stopniu podważa cel, jaki postawiłem sobie do realizacji w niniejszych rozważaniach, mianowicie: uzasadnienie normatywności mechanizmów predykcyjnych. Jeżeli nie ma żadnych norm w świecie, to mechanizmy generowane przez bayesowski model generatywny również normatywne nie są. Także mówienie o normatywności modelowania bayesowskiego w kognitywistyce czy psychologii ostatecznie można uznać za przejaw instrumentalizmu albo – szerzej ujmując – jakiejś formy antyrealizmu[19].

Sądzę jednak, że wnioski Daviesa są w pewnym stopniu przedwczesne. Twierdzę, że część funkcji biologicznych nie jest normatywna, a część jest. To znaczy istnieje

[18] Czyni tak np. Bielecka (2018), proponując ujęcie hybrydowe, które łączy teleosemantykę z koncepcją systemową.

[19] Przez antyrealizm w najbardziej ogólnym sensie rozumiem tutaj taki pogląd, który odrzuca bądź zawiesza tezę realistyczną.

taka klasa funkcji, dla jakich możliwość dysfunkcji jest konstytutywna, a nie tylko przygodna. Innymi słowy, mogą istnieć takie funkcje, jakich normatywność da się uzasadnić systemowo, czyli że normatywność jest ich cechą definicyjną i – co ważniejsze – konstytutywną. Mówiąc o takich funkcjach, będę miał na myśli funkcję reprezentowania oraz funkcję predykcyjną.

4.3.4. Funkcja reprezentowania

Pojęcie funkcji reprezentowania jest istotne dla wielu koncepcji z zakresu filozofii biologii (por. m.in. Bickhard, 2014; Dretske, 1981, 2014). Obecnie skoncentruję się na propozycji Millikan (2011, 2014), ponieważ, jak już wiemy, wiąże ona pojęcie funkcji z pojęciem normatywności. Oględnie rzecz ujmując, funkcją właściwą systemów percepcyjnych i poznawczych jest funkcja reprezentowania. Jest to funkcja właściwa, która mówi nam, kiedy ma się do czynienia z jej dysfunkcją. Dysfunkcja polega na niespełnianiu funkcji właściwej, czyli w tym przypadku na niewłaściwym, niepoprawnym czy ostatecznie błędnym reprezentowaniu. Według Bieleckiej możliwość pojawienia się błędnej reprezentacji stanowi o tym, że funkcja reprezentowania jest jakąś funkcją, która spełnia lub może spełniać określone warunki Normalne. Oznacza to, że błędne reprezentowanie jest istotnym i konstytutywnym elementem normalnego funkcjonowania organizmu, ponieważ „z perspektywy ewolucyjnej bardziej opłacalne jest nadmierne, a więc błędne reprezentowanie pewnych cech otoczenia, niż niereprezentowanie ich w ogóle" (Bielecka, 2018, s. 185). To, co jest istotne dla tego ujęcia, to (1) stwierdzenie, że nie da się wyjaśnić możliwości błędnego reprezentowania bez odniesienia do historii ewolucyjnej danego organu lub jego części; oraz (2) że dla wyjaśnienia funkcji reprezentowania przez określone organy bądź szerzej – organizmy konieczne jest wyjaśnienie dysfunkcji czy określenie normatywnych warunków właściwych dla funkcji reprezentowania.

To, co stanowi o atrakcyjności propozycji Millikan, jednocześnie określa jej problematyczność. Po pierwsze jeżeli przyjmie się, że wyjaśnienie dysfunkcji systemów reprezentacyjnych[20] opiera się na odwołaniu do historii selekcji naturalnych ich poprzedników, to pojawia się zarzut sformułowany m.in. przez Daviesa. Po drugie zaś ujęcie Millikan oferuje opis normatywnej funkcji reprezentowania tylko w perspektywie gatunkowej, a nie indywidualnej. W praktyce wiąże się to z tym, że możemy podać jedynie minimalne warunki, jakie dany system reprezentacyjny powinien spełniać, aby mógł poprawnie lub błędnie reprezentować, niemniej nie możemy wyjaśnić, dlaczego dany organizm czy system zachował się w taki lub inny sposób bądź wytworzył taką, a nie inną reprezentację. Chodzi o to, że mówienie

[20] Przez takie systemy będę najogólniej rozumiał te, które spełniają funkcję reprezentowania.

o normatywności w teleosemantyce służy wyłącznie uzasadnieniu takiego, a nie innego rodzaju opisu systemów biologicznych i sztucznych. Normatywność jest zatem tutaj pojęciem opisowym, a nie wyjaśniającym, co w praktyce oznacza, że gdybyśmy chcieli wykorzystać pojęcie funkcji właściwej do wyjaśnienia normatywności bayesowskiego modelu generatywnego i jego mechanizmów predykcyjnych, to powinniśmy to wyjaśnienie sprowadzić do konstatacji, że wyjaśnienie modelu polega na wykazaniu warunków, które musi on spełniać, aby mógł (przykładowo) błędnie lub niepoprawnie minimalizować błędy predykcyjne. Takie ujęcie jest jednak niesatysfakcjonujące, bowiem model ten z definicji działa w ten sposób, że za każdym razem jest on podatny na błędną reprezentację. Mówiąc technicznym językiem PP: statystyczny model generatywny, określając prawdopodobieństwa wstępne generowanych przez siebie predykcji, nieustannie porównuje je z przychodzącymi danymi i w ten sposób aktualizuje swoje wewnętrzne parametry, jednocześnie określając nowe rozkłady prawdopodobieństwa dla określonych wartości zmiennych, które przyjmuje. Cały ten proces polega na możliwie dokładnym zbliżeniu do siebie rozkładów prawdopodobieństwa parametrów wewnętrznych i przypisanych rozkładów prawdopodobieństwa zewnętrznym stanom środowiska. Rozbieżność pomiędzy oboma rozkładami jest określona w PP mianem „dywergencji Kullbacka-Leiblera" (dywergencja KL). Dotyczy ona bowiem różnicy pomiędzy rzeczywistymi danymi a modelem teoretycznym (tj. wewnętrznymi parametrami modelu generatywnego) (por. Bogacz, 2017; Kiefer, Hohwy, 2018). Oznacza to, że błędne reprezentowanie jest wpisane w samą naturę modelu generatywnego.

Model może reprezentować tylko wtedy, gdy jest możliwe, aby reprezentował błędnie (Bielecka, 2018). Takie ujęcie jednak jest rozstrzygnięciem pojęciowym, ponieważ mówienie o reprezentowaniu zakłada z definicji możliwość błędnej reprezentacji: jeżeli *X* może reprezentować błędnie, to *X* reprezentuje. Oznacza to, że nie mamy tutaj do czynienia z wyjaśnieniem w sensie przyczynowym, a raczej z analizą pojęciową[21]. Jest ona oczywiście potrzebna w nauce, a szczególnie w filozofii, niemniej takie przypisanie możliwości błędnej reprezentacji (która ma stanowić o normatywności danej funkcji czy mechanizmu) danym narządom, systemom czy mechanizmom ma charakter epistemiczny, bowiem jest zrelatywizowane do opisu formułowanego przez zewnętrznego obserwatora w odniesieniu do jego interesów badawczych, nie zaś charakter ontyczny, czyli wywiedziony ze struktury danego organizmu, systemu czy mechanizmu. Twierdzę zatem, że powinno się raczej podać warunki przyczynowe, jakie model (czy szerzej – organizm bądź system) musi spełniać, aby na podstawie probabilistycznych reprezentacji (czyli mniej lub bardziej prawdopodobnych) mógł działać w środowisku w sposób adaptacyjny i taki, jaki

[21] Innymi słowy: atrybucja dysfunkcji zakłada atrybucję funkcji. W tym sensie funkcja i dysfunkcja definiują się nawzajem. Problematyczne jest zatem stwierdzenie, że funkcja wyjaśnia możliwość dysfunkcji. Dziękuję Pawłowi Gładziejewskiemu za zwrócenie mi uwagę na ten aspekt problemu funkcji.

przynosi sukcesy poznawcze i pozapoznawcze określonemu indywiduum, które jest „uzbrojone" w model generatywny.

Podanie normatywnych warunków umożliwiających błędną reprezentację to zaledwie część pracy, którą należy wykonać. Może być przecież tak, że organizm prawidłowo reprezentuje świat, a mimo wszystko jego działania są nieskuteczne. Przykładowo: mogę mieć dobrą reprezentację drzewa, na które mógłbym wejść, aby zdobyć pożywienie, a nieprawidłowo oszacuję ryzyko wejścia na drzewo, w efekcie czego mój zamiar się nie powiedzie. Zatem sama funkcja reprezentowania nie wystarczy, aby wyjaśnić normatywność mechanizmu predykcyjnego. Istotna jest również jego relacja do określonych zmiennych ze środowiska. Innymi słowy, sam fakt tego, że model generatywny jest reprezentacyjny, nie tłumaczy, dlaczego może on skutecznie bądź nie przewodzić działaniami organizmu.

Pewną odpowiedź na powyższe problemy sformułował Bickhard w swojej wersji interakcjonizmu i koncepcji funkcji normatywnych. Omówię teraz jego propozycję.

4.3.4.1. Funkcje w ujęciu Bickharda

Bickhard, podobnie jak Millikan, twierdzi, że aby móc wyjaśnić funkcję reprezentacji, należy odnieść się do możliwości popełnienia błędu reprezentacyjnego przez dany system. Innymi słowy, wyjaśnienie możliwości reprezentowania czegoś opiera się na normatywnej funkcji reprezentowania, czyli tej, która pozwala wyjaśnić możliwość pojawienia się dysfunkcji (Bickhard, 2014, s. 139). Bickhard zatem przyjmuje za Millikan, że funkcje biologiczne są jednocześnie funkcjami normatywnymi. Odrzuca on jednak ich etiologiczne uzasadnienie, proponując swoje oparte na pojęciach interakcji i emergencji. Zanim jednak je omówię, konieczne jest przedstawienie jego propozycji teoretycznej w interesujących nas aspektach[22].

Kluczowe dla koncepcji Bickharda są pojęcia autonomii i samoorganizacji. Pojęcia te naturalnie kojarzą się z teoriami autopojetycznymi. Nie można jednak ujęcia Bickharda traktować jako rozwinięcia projektu Maturany i Vareli (por. 1980, 1992). Przypomnijmy: ich zdaniem ciągła tendencja do samoorganizacji jest charakterystyczna i fundamentalna dla wszystkich układów żywych. Z tego powodu określają je oni mianem „autopojetycznych", tj. takich, jakie nieustannie podtrzymują się przy życiu dzięki ciągłej realizacji procesów życiowych. Procesy te są niezależne informacyjnie od środowiska. Według zwolenników tego ujęcia autonomię należy rozumieć w kategoriach autopojetycznych. Bickhard jednak proponuje, aby autonomię ujmować w kategoriach termodynamicznych: organizm

[22] Należy podkreślić, że nie będę omawiał istotnej dla Bickharda ontologii procesu i zagadnienia emergencji. Sam interakcjonizm omówię jedynie w tym stopniu, który jest kluczowy dla zrozumienia jego koncepcji normatywności.

żyje tak długo, jak długo pozostaje w nierównowadze energetycznej z otoczeniem. Takie ujęcie zbliża propozycję Bickharda do pozycji zajmowanej przez Fristona. Na pytanie o to, czy tak rzeczywiście jest, odpowiem niebawem. Bickhard powiada, że systemy autopojetyczne, w tym organizmy żywe[23], realizują procesy, które w interakcji z otoczeniem wykonują taką pracę, która ma za zadanie dostarczenie energii do systemu. Paradygmatycznym przykładem takiego systemu jest żywa komórka, która wykorzystując procesy metaboliczne, przekształca energię i materiały ze środowiska w energię chemiczną i molekuły organiczne, które są konieczne dla procesów podtrzymujących komórkę przy życiu. W tym kontekście Bickhard wprowadza pojęcie funkcji normatywnej. Powiada: „stabilność procesów dalekich od równowagi wymaga ich utrzymywania, a takie utrzymywanie jest funkcjonalne względem stabilności tego systemu – *sprawuje ono pewną funkcję*, o ile przyczynia się do tej stabilności" (Bickhard, 2014, s. 121). Takie pojęcie funkcji określa on mianem „podstawowego". Tak rozumiana funkcja przyczynia się do utrzymania stabilności danego procesu, tym samym wspomagając utrzymywanie istnienia organizmu. Jest ona zatem normatywna dla tej stabilności. Jest ona również normatywna, ponieważ utrzymując stabilność danego procesu, przyczynia się do realizacji i utrzymania stabilności innych procesów (Christensen, Bickhard, 2002, s. 4). Takie rozumienie normatywności pozwala Bickhardowi odróżnić organizację przyczynowo-skutkową i organizację normatywną. Normy są przezeń traktowane jako własności wewnętrznie organizacyjne, ponieważ są one cechami relacji zachodzącej pomiędzy częściami systemu a globalną organizacją tego systemu. W ten sposób tak rozumiane normy określają możliwe powiązania oraz role dla określonych części systemu (tzw. współzależność procesowa). Ograniczają one zatem możliwe relacje i zachowania systemu. Tam gdzie pojawiają się takie normatywne ograniczenia, zachowana jest globalna organizacja systemu. Poza jego granicami jest ona zakłócona. Normy są zatem kluczowym aspektem organizacji systemu, a nie czymś od niego odrębnym (Christensen, Bickhard, w przygotowaniu, s. 4). Christensen i Bickhard podkreślają, że współzależność procesowa jest centralnym zagadnieniem w modelowaniu i zrozumieniu funkcji normatywnej. Relacje współzależności procesowej określają naturę odpowiednich systemów. Ich sukces lub porażka determinują sukces bądź porażkę systemu jako faktycznego i autonomicznego. W takim ujęciu części i procesy spełniają funkcje normatywne, o ile spełniają one również wymagania innych procesów w systemie. Są one zatem normatywne, ponieważ realizują normatywne funkcje (Christensen, Bickhard, w przygotowaniu, s. 16; por. także Garson, 2013).

Podsumujmy nasze dotychczasowe ustalenia. Zdaniem Bickharda funkcja jest normatywna nie tylko dlatego, że pozwala wyjaśnić sytuacje, w których staje się ona

[23] Za systemy autonomiczne Bickhard uznaje także gatunki, niektóre zbiorowiska biologiczne oraz systemy społeczno-kulturowe (np. przedsiębiorstwa, miasta) czy regiony geopolityczne (por. Christensen, Bickhard, 2002).

dysfunkcjonalna, ale przede wszystkim ze względu na określone role przyczynowe, które odgrywa w organizacji i utrzymaniu systemu żywego w nierównowadze termodynamicznej z otoczeniem. Mówiąc precyzyjniej: funkcja jest normatywna, wtedy, gdy kształtuje i ogranicza możliwe relacje, procesy i zachowania danego systemu. Procesy i funkcje są normatywne, a nie tylko przyczynowe, ponieważ utrzymują one stabilność systemu. Nie jest jednak tak, że zawdzięczają swoją normatywność tej stabilności. To raczej to, co przyczynia się do utrzymania danej stabilności, jest dla tej stabilności normatywne. „I takie przyczynianie jest decydujące dla tego, jak toczy się bieg spraw tego świata" (Bickhard, 2014, s. 122, przyp. 28). Przykładowo funkcją serca jest pompowanie krwi i jest to funkcja normatywna, ponieważ (1) umożliwia ona inne procesy (np. dostarczanie tlenu do mózgu) oraz (2) przyczynia się w sposób konieczny do przetrwania systemu. Nie mamy tu zatem do czynienia tylko z funkcją przyczynową, ale także z funkcją konieczną, tj. normatywną, która jest emergentna względem określonych mocy przyczynowych (por. Bickhard, 2002, 2003, 2014, s. 118–121)[24]. Jest normatywna, bowiem jest konieczna dla utrzymania organizmu przy życiu. Funkcja serca do wydania efektów akustycznych jest tylko funkcją przyczynową, nie jest zaś funkcją normatywną, ponieważ nie przyczynia się ani do realizacji innych procesów, ani do stabilności i przetrwania organizmu. Przykładowo: funkcja serca do pompowania krwi sprawia, że pełniona jest funkcja dostarczania tlenu do mózgu. Oznacza to, że ta pierwsza funkcja nie może być realizowana, o ile nie przyczyni się ona do realizacji funkcji drugiej. Zatem zależność pomiędzy obiema funkcjami jest normatywna.

Należy zatem stwierdzić, że normatywności w ujęciu Bickharda nie można potraktować jedynie jako pewnego predykatu, który przypisuje się funkcjom w celu wyjaśnienia możliwych nieprawidłowości w ich działaniu. Jak pamiętamy, stanowisko Millikan i innych zwolenników teleosemantyki jest skazane na taki zarzut, co dobitnie wykazał Davies. W przypadku propozycji Bickharda normatywność jest naturalną cechą funkcji, ponieważ jeżeli odgrywają one określone role przyczynowe, to są normatywne[25]. Koncepcja Millikan, o czym już wspomniałem, wymogi normatywny i eksplanacyjny spełnia jedynie w sensie słabym. Odwołuje się bowiem do słownika normatywnego, ale nie wykazuje zależności normatywnych. Inaczej jest w przypadku Bickharda. W jego ujęciu o funkcjach możemy orzekać, że są normatywne, ponieważ dzięki nim są możliwe określone zależności normatywne.

Powróćmy na chwilę do podobieństwa modelu Bickharda z modelami rozwijanymi przez Fristona. Obaj badacze podkreślają konieczność uwzględnienia relacji energetycznych systemu z otoczeniem. Obaj stwierdzają, że żywy organizm, aby mógł przetrwać, musi unikać równowagi energetycznej. I tutaj, jak sądzę, istotne podobieństwa pomiędzy propozycjami tych autorów się kończą. Bickhard w prze-

[24] Na temat pojęcia emergencji por. Poczobut, 2009.
[25] Takie ujęcie pozwala Bickhardowi na powiązanie wymogu eksplanacyjnego z normatywnym.

ciwieństwie do Fristona nie poszukuje jednej ogólnej, eksplanacyjnej zasady na wzór FEP, która wyjaśniałaby funkcjonowanie i działanie organizmów żywych. Friston, przypomnijmy, za mechanizmy normatywne uważa te, które są homeostatyczne. Te zaś należy rozumieć jako mechanizmy teleologiczne. Dzięki tej ich własności można je określać mianem „normatywnych": działają one poprawnie, jeżeli spełniają cel, ze względu na który wyewoluowały, lub działają niepoprawnie, jeżeli tego celu nie realizują. Oznacza to, że o normatywności określonej funkcji decyduje jej rola w utrzymaniu homeostazy, a nie przyczynowa rola w funkcjonowaniu systemu jako takiego. Jest to poważne ograniczenie, ponieważ sprowadza ono mówienie o funkcjach normatywnych tylko do tych, które można powiązać z mechanizmami homeostatycznymi. Godfrey-Smith przekonująco wykazał, że trzeba zachować szczególną ostrożność przy stwierdzaniu, że dany mechanizm służy homeostazie (Godfrey-Smith, 1996, s. 79; zob. podrozdział *Normatywność homeostazy*). W praktyce ujęcie Fristona wiąże się z tym, że o wielu funkcjach (np. motorycznych, percepcyjnych czy poznawczych) nie można stwierdzić, że są normatywne. Jest to jednak mocno problematyczne. Trudno jest bowiem przykładowo wyobrazić sobie sytuację, w której o procesie decyzyjnym prowadzącym do pojęcia określonych działań nie można będzie powiedzieć, że jest normatywny. Model Bickharda na takie orzekanie pozwala. Wiąże on bowiem normatywność funkcji z ogólną architekturą systemu i z relacjami, które nią rządzą.

Jak dotąd udało się wykazać, że normatywność w ujęciu Bickharda dotyczy struktury i funkcjonalności w obrębie systemu. Wydaje się, że proponowane ujęcie jest odporne na zarzut, który sformułowałem wobec koncepcji Millikan, czyli przypisywania normatywności funkcjom w odniesieniu do typów biologicznych, a dopiero wtórnie egzemplarzom tych typów (por. Millikan, 2011, s. 121–123). Wiąże się ono bowiem z indywidualną strukturą określonego osobnika. Dlaczego? Ponieważ Bickhard wykazuje, że normatywność jest wewnętrzną własnością określonych funkcji oraz relacji, a nie skutkiem doboru naturalnego[26]. Z tego też powodu wydaje się, że jego ujęcie ma charakter ontyczny, a nie epistemiczny, czyli że spełnia wymogi eksplanacyjny i normatywny w sensie mocnym. Pojawia się jednak następujący problem: określona organizacja funkcjonalna i strukturalna nie jest wynikiem indywidualnego rozwoju. Jest ona efektem długotrwałych procesów ewolucyjnych. Można zatem stwierdzić, że ujęcie Bickharda, mimo że zrywa z podejściem etiologicznym, to ostatecznie daje się do niego sprowadzić, ponieważ na koniec i tak należy postawić pytanie o genezę funkcji. Można na to odpowiedzieć, twierdząc, że normatywność funkcji nie jest zależna od ich historii ewolucyjnej, bowiem istnieje fundamentalna różnica pomiędzy posiadaniem da-

[26] Nie chodzi o to, że uzasadnienie etologiczne jest fałszywe, bo nie jest, tylko o to, że nie wyjaśnia ono dobrze, dlaczego określone funkcje są normatywne. Wydaje się, że jest to uzasadnienie zewnętrzne i zależne od sposobu, w jaki wyjaśnia się naturę organizmów żywych.

nej funkcji (co ostatecznie wyjaśnia nam ich etiologia) a jej pełnieniem. Bickhard (2014, s. 125) zauważa:

> W przypadku etiologicznym sprawowanie funkcji występuje tylko wtedy, gdy coś mającego funkcję faktycznie prowadzi do konsekwencji spełniających te funkcje. Nerka sprawuje funkcję filtrowania krwi, kiedy faktycznie ją filtruje, a czyni tak, ponieważ ma funkcję filtrowania krwi. Należy zauważyć, że w tej koncepcji żadna funkcja nie może być sprawowana, o ile coś nie *ma* tej funkcji. Nieco trudniejsze jest zrozumienie posiadania funkcji w kategoriach sprawowania funkcji […], ale jedna zaleta wyboru „sprawowania funkcji" jako przedmiotu eksplikacji jest oczywista: funkcje mogą być sprawowane, nawet jeśli nic nie ma tych funkcji.

Oznacza to, że system, aby mógł przetrwać i realizować swoje zadania, musi pełnić określone funkcje. W takim ujęciu sztuczna zastawka serca pełni tę samą funkcję normatywną, co zastawka naturalna. Nie da się jednak jej normatywności sprowadzić do jakiegoś wyjaśnienia etiologicznego.

Bickhard w swoim ujęciu nie tylko przekracza ograniczenia koncepcji teleosemantycznej, ale także innych koncepcji funkcji, którym można zarzucić internalizm. Mianowicie: funkcje oraz ich normatywność są wyjaśniane poprzez odniesienie do wewnętrznej architektury systemu oraz ról, jakie odgrywają w tej architekturze. Wydaje się jednak, że tak rozumiana normatywność nie jest regulowana tylko przez czynniki wewnętrzne, ale także zewnętrzne. Dużą świadomość tego faktu zdaje się mieć Millikan, jednak ze względu na założenia swojej teorii nie wykorzystuje ona w pełni potencjału tego odkrycia. Twierdzi bowiem, że celem funkcjonowania systemów percepcyjnych i poznawczych jest wytworzenie pewnych struktur relacyjnych ze środowiskiem (Millikan, 2011, s. 122–123). Niestety, w jej ujęciu wpływ tego środowiska jest analizowany tylko w perspektywie historycznej. Nie ma zatem mowy o takim ujęciu, jakie odnosi się do faktycznego usytuowania organizmu w *tu* i *teraz* określonej sytuacji środowiskowej. Propozycja Bickharda pozwala mówić w pewnym stopniu o takiej możliwości. Jego zdaniem pełnienie określonej funkcji jest zależne od wielu różnych warunków, w tym także środowiskowych. Tworzenie lub utrzymywanie tych warunków samo w sobie pełni określone funkcje, bowiem samo ono pełni kolejną funkcję wspierania pełnienia kolejnych funkcji itd. (Bickhard, 2014, s. 125–126). Oznacza to, że mamy tutaj do czynienia z czymś, co roboczo można określić jako „normatywność wewnętrzną" (systemową) i „zewnętrzną" (środowiskową). Jest to istotne odkrycie, bowiem umożliwi sformułowanie nietrywialnej koncepcji biologicznych mechanizmów normatywnych w ogóle (nie tylko normatywnych mechanizmów predykcyjnych).

4.3.4.2. Normatywność reprezentacji

Przypomnijmy: organizm może przetrwać, jeżeli unika równowagi energetycznej ze środowiskiem. Oznacza to, że normatywność wielu jego mechanizmów można uzasadnić jedynie poprzez odniesienie do wewnętrznej struktury danego systemu. Są one normatywne z powodu funkcji, którą sprawują, oraz dlatego, że współprzyczyniają się do realizacji innych procesów w tym systemie. Bickhard zwraca jednak uwagę na fakt, że nie da się w ten sposób wyjaśnić normatywności wszystkich mechanizmów i realizowanych przez nie funkcji. Jego zdaniem system za sprawą niektórych mechanizmów oddziałuje na środowisko i ze środowiskiem. Przykładem mechanizmów, które mają wykraczać poza internalistyczne wyjaśnienie, ponieważ ze swej istoty wchodzą w interakcje ze środowiskiem, są mechanizmy reprezentacyjne oraz, za czym będę argumentował, mechanizmy predykcyjne[27]. Kluczowa dla mechanizmów reprezentacyjnych jest kwestia prawdziwości. Generowane przez nie reprezentacje mają wartość logiczną prawdy lub fałszu. Normatywność reprezentacji odnosi się do faktu, że mogą być one prawdziwe lub fałszywe (Bickhard, 2003, 2015a, 2015b). Owa wartość reprezentacji jest kluczowa dla organizmu z pragmatycznych powodów. Zwiększa bowiem lub zmniejsza szansę na sukces bądź porażkę działań organizmu w środowisku. Owa normatywna własność reprezentacji wyłania się z procesów (Bickhard używa pojęcia emergencji), które są kluczowe dla przeżycia organizmu. Dla owej emergencji nie są kluczowe kwestie etiologiczne, co nieustannie jest podkreślane, ale interakcje, w jakie organizm wchodzi ze swoim otoczeniem. Nie chodzi tutaj zatem o wyjaśnienie emergencji zjawisk normatywnych ze zjawisk nienormatywnych, ale o pokazanie, w jaki sposób zjawiska normatywne przyczyniają się do zachowania autonomii przez system (Bickhard, 2014, 2015a, s. 226). Jego zdaniem sama normatywność nie potrzebuje uzasadnienia, ponieważ fakt autonomii organizmów żywych ją uzasadnia. Należy jednak wyjaśnić sposób, w jaki zjawiska normatywne wspomagają autonomię.

Co wynika ze stwierdzenia, że reprezentacje są normatywne w sensie bycia prawdziwymi lub fałszywymi? Z jednej strony wiadomo już, że mechanizmy je wytwarzające są normowane przez ogólną architekturę systemu oraz relacje, w jakie kolejne procesy ze sobą wchodzą. Te ostatnie można wyjaśnić poprzez odniesienie do tego, co określiłem jako „normatywność wewnętrzną". W przypadku funkcji reprezentowania takie ujęcie jest niewystarczające, ponieważ nie da się orzec wartości logicznej danej reprezentacji, jeżeli nie odniesie się jej do środowiska. Nie chodzi tutaj jednak o prawdę w sensie korespondencji, czyli o klasycznie pojmowaną prawdziwość, ale o prawdę w ujęciu pragmatycznym, której kryterium jest sukces

[27] Nie wnikam teraz w kwestię tego, czy niektóry mechanizmy predykcyjne są jednocześnie mechanizmami reprezentacyjnymi jak przykładowo twierdzą Gładziejewski (2016) czy Hohwy oraz Kiefer (2018). Problem ten pozostawiam na boku moich analiz.

bądź porażka organizmu działającego w środowisku (Bickhard, 2014, s. 140–141, 2015a, s. 226). Bickhard (2014, s. 141) stwierdza, że:

> aktywności antycypacyjne mają wartość logiczną. Jeżeli nie przebiegają tak, jak antycypowano, są fałszywe i *sfalsyfikowane dla organizmu*. To znaczy: jeśli wewnętrzna gotowość do angażowania się w interakcję zostanie naruszona, jeżeli rzeczywisty przebieg interakcji wykroczy poza granice tego, co jest funkcjonalnie antycypowane, to ta antycypacja jest sfalsyfikowana – dla samego organizmu.

Należy zatem powiedzieć, że funkcja reprezentowania jest zrelatywizowana do aktywności systemu względem warunków środowiskowych. Kluczowa jest również możliwość wykrycia błędu (np. błędnej reprezentacji) przez system bez uwzględniania zewnętrznych kryteriów. Gwarantem tej możliwości są udane bądź nieudane interakcje (Bickhard, 2014, s. 130–131)[28]. W pewnym sensie wydaje się, że omawiany model jest internalistyczny, bowiem prawdziwość bądź fałszywość reprezentacji wiąże się z utrzymaniem autonomii systemu, a nie odniesieniem do faktycznych własności środowiska. Bickhard jednak przyjmuje, że środowisko w jakimś stopniu wspiera określone interakcje, bowiem reprezentacje mają swoją treść (Bickhard, 2015a, s. 226). Wydaje się jednak, że jest to warunek dość słaby, bowiem w jego koncepcji nie treść decyduje o wartości logicznej reprezentacji, ale względy pragmatyczne, które koniec końców można wyjaśnić poprzez procesy i mechanizmy wewnętrzne, których normatywność jest, o czym już pisałem, rozumiana internalistycznie (por. Bickhard, 2015b, s. 4). Twierdzę, że pełne wyjaśnienie normatywności takich mechanizmów jak mechanizmy reprezentacyjne wymaga odwołania się w tym samym stopniu do czynników zewnętrznych, jak i wewnętrznych. Tym sposobem wykaże się istnienie nie tylko przyczynowych zależności, ale także tych, które można określić jako „motywacyjne" czy „zobowiązujące", czyli normatywne w mocnym sensie tego słowa. Aby móc uzasadnić prezentowane tu tezy, przejdę do dyskusji z propozycją Bickharda.

4.3.4.3. Zarzuty Bickharda wobec przetwarzania predykcyjnego

Bickhard podkreśla, że PP łączy wiele podobieństw z jego własnym ujęciem z powodu podkreślania istotnego znaczenia w percepcji zorientowanych na przyszłość predykcji (2015b, 2016). Nie będę przedstawiał wszystkich aspektów jego krytyki, bowiem nie są one istotne dla podejmowanych tutaj analiz. Ograniczę się zatem do tych jej punktów, które są ważne dla prowadzonych tu badań. Na

[28] Problemem tym nie będę się zajmował, ponieważ znacznie wykracza on poza ramy przyjęte dla tych rozważań. Jest to jednak zagadnienie kluczowe dla wyjaśnienia pewnych aspektów normatywności reprezentacji także w ramach PP. W tym miejscu zadowalam się konstatacją, że reprezentacje są normatywne, ponieważ mogą być prawdziwe lub fałszywe bądź błędne.

początek należy powiedzieć, że to, co jest charakterystyczne dla jego interpretacji, to powiązanie PP z fristonowską FEP. Argumentowałem wcześniej, że takie powiązanie nie jest w pełni uzasadnione z powodu eksplanacyjnej słabości propozycji Fristona, niemniej zgadzam się z Bickhardem, że wiele rozwiązań przyjętych w ramach badań nad FEP jest interesujących i ważnych dla PP (por. podrozdział *Zastosowanie zasady energii swobodnej w przetwarzaniu predykcyjnym*). I tak Bickhard zauważa, co również było już przeze mnie wielokrotnie podkreślane, że problem normatywności w PP jest praktycznie nieobecny i niesie to ze sobą określone konsekwencje. Po pierwsze, jego zdaniem, w ramach PP utrzymuje się, że wyjaśnienie możliwości minimalizacji błędów predykcyjnych lub mówiąc innym językiem – zaskoczenia nie wymaga powoływania się na żadne pojęcia normatywne, takie jak „cele", „preferencje" czy „wartości". Nie jest to do końca prawda, bowiem Friston wielokrotnie powołuje się na tego typu pojęcia, niemniej rozumie je w kategoriach czysto statystycznych, co w efekcie czyni je pojęciami, których normatywne znaczenie jest niejasne (por. Friston, Adams, Montague, 2012; Friston i in., 2015b; podrozdziały *Wnioskowanie aktywne – Uwagi wobec ramy aktywnego wnioskowania*)[29]. Po drugie jeżeli podstawowym kryterium dla podjęcia odpowiedniego działania jest skuteczność predykcji w minimalizowaniu błędu predykcyjnego, to dlaczego organizm nie chce po prostu udać się do ciemnego pokoju i tam zostać (por. Clark, 2013b; Klein, 2018; podrozdziały *Problem ciemnego pokoju*, *Problem ciemnego pokoju a mechanizmy normatywne*)? Takie ujęcie wydaje się kontrintuicyjne, bowiem teoretycznie nikt nie chciałby spędzać swojego życia w ciemnym, pozbawionym bodźców pomieszczeniu. Łatwo można odpowiedzieć na ten zarzut w duchu Fristona, twierdząc, że przykładowo organizm oczekuje, że pokój, do którego ma wejść, będzie oświetlony, i z tego powodu włącza światło, aby swoje oczekiwania zrealizować (por. Friston i in., 2012a). Oznacza to, że wszelkie normatywne założenia miałyby być wpisane w wysoko poziomowe predykcje (*hyperpriors*), czyli przekonania[30]. Możemy sobie jednak bez trudu wyobrazić sytuację, w której ktoś z pewnych powodów postanawia wejść do ciemnego pokoju i tam pozostać, albo przykładowo odmawia on sobie przez dłuższy czas pokarmów, co w żaden sposób nie odpowiada podstawowym potrzebom organizmu zrelatywizowanym do przetrwania w zmiennym i niestabilnym otoczeniu. Jeszcze innym przykładem jest sytuacja, w której ktoś rozmyślnie zwiększa niepewność, podejmując określone działania, np. zwiększające prawdopodobieństwo jego śmierci. Jak jest

[29] Jest ono niejasne, gdyż proponowana przez niego teoria oparta na FEP ma być normatywną teorią działania i percepcji, ponieważ zapewnia dobrze zdefiniowaną funkcję obiektywną (wariacyjna energia swobodna), która jest optymalizowana zarówno przez działanie, jak i percepcję (Schwartenbeck i in., 2013, s. 1; por. rozdział *Mechanizmy normatywne a działania w przetwarzaniu predykcyjnym*).

[30] Przez przekonania rozumiem określone rozkłady prawdopodobieństwa dotyczące nieznanych stanów świata, generowane przez wielopoziomowy model generatywny.

zatem możliwy konflikt pomiędzy jednymi oczekiwaniami a niesprowadzalnymi do konieczności przetrwania potrzebami danego osobnika?

Inny ważny zarzut Bickharda dotyczy minimalizacji niepewności. Sama jej konieczność nie wystarcza, aby mówić o normatywności, należy ją bowiem traktować jako „domniemaną konsekwencję wcześniejszej selekcji ewolucyjnej". Oznacza to, że funkcjonowanie organizmu tłumaczy się na podstawie istnienia faktycznych i przypadkowych warunków, które minimalizują rozbieżności pomiędzy wewnętrznie generowanymi predykcjami a sygnałami pochodzącymi z wejść zmysłowych (Bickhard, 2016, s. 264). W tym sensie trudno jest wyjaśnić, w jaki sposób i dlaczego organizm szuka wartości lub unika szkód. Należałoby stwierdzić, że czyni to ze względu na ewolucję lub trening, co znowu nas prowadzi do problemu ciemnego pokoju. Friston, Daunizeau i Kiebel (2009) twierdzą, że wysokopoziomowe oczekiwania są „wbudowane" w organizację całego organizmu. Jednak przyjęcie takiej hipotezy nie pozwala ostatecznie wyjaśnić normatywnej różnicy pomiędzy udanym działaniem prowadzącym do jakiegoś sukcesu a przypadkowym zachowaniem lub jakimś błędem albo pomyłką, ponieważ z punktu widzenia organizacji systemu wszystkie takie procesy są po prostu przyczynowe. Jest to mocny zarzut, jednak, za czym już argumentowałem w tej pracy, w PP mechanizmy predykcyjne i generowane przezeń predykcje są w pewnym sensie normatywne ze względu na ich kluczową rolę w minimalizowaniu błędów predykcyjnych. Jednakże, jak podkreśla Bickhard i z czym trudno się nie zgodzić, rola ta, jeżeli nie wykaże się inaczej, dotyczy tylko faktycznych i przyczynowych procesów. Kluczową zatem kwestią jest odróżnienie takich relacji pomiędzy predykcjami a działaniami, jakie są nie tylko przyczynowe, ale także normatywne. Przyjmując podejście Bickharda, należy stwierdzić, że aby móc wyjaśnić normatywność predykcji, musimy pokazać, jak ta normatywność wyłania się z naturalnej organizacji danego organizmu. Nie oznacza to jednak, że powinniśmy ograniczyć się tylko do wewnętrznej architektury modelu generatywnego, ale należy również wziąć pod uwagę relacje, w jakie organizm, „uzbrojony" w model generatywny, wchodzi ze swoim środowiskiem (por. Piekarski, 2019, s. 2). To wymaga jednak rozwinięcia uzasadnienia, które będzie celem moich dalszych analiz.

4.4. Model normatywności dla przetwarzania predykcyjnego

Na początek należy wprowadzić pewne podstawowe rozróżnienia pojęciowe. W różnych miejscach mówię o normatywności mechanizmów, funkcji i predykcji. O relacji pomiędzy funkcjami i mechanizmami już wielokrotnie wspomniałem. Tutaj zatem ograniczę się do stwierdzenia, że dany mechanizm jest normatywny

wtedy, gdy realizuje on określoną funkcję normatywną lub określone funkcje normatywne. Funkcja zaś jest normatywna wtedy, gdy jednocześnie (1) pozwala wyjaśnić dysfunkcję danego mechanizmu; (2) przyczynia się do utrzymania stabilności organizmu w ten sposób, że kształtuje i ogranicza możliwe relacje, procesy i zachowania danego systemu (w taki sposób, że to, co przyczynia się do utrzymania danej stabilności, jest dla tej stabilności normatywne); oraz dodatkowo – w odniesieniu do funkcji reprezentowania i funkcji predykcyjnej – (3) gdy można za jej pomocą wyjaśnić możliwość przypisywania wartości logicznych określonym reprezentacjom/predykcjom, w efekcie czego pozwala ona wyjaśnić możliwość błędu oraz uzasadnić, dlaczego określona reprezentacja/predykcja może kierować działaniami[31]. Dla moich analiz kluczowa jest trzecia własność funkcji normatywnych, której jak dotąd nie omówiłem dokładnie.

Przez pojęcie predykcji, najogólniej mówiąc, rozumiem przewidywanie probabilistyczne, które wyraża statystyczny związek zachodzący pomiędzy określonymi wartościami pewnych zmiennych (np. zależność pomiędzy ciśnieniem atmosferycznym a opadami). Tak rozumiana predykcja jest wnioskiem wnioskowania abdukcyjnego, tj. wnioskowania do najlepszego wyjaśnienia, co w efekcie oznacza, że dla jej wyjaśnienia konieczne jest założenie istnienia modelu generatywnego rozumianego jako wielopoziomowa sieć bayesowska (por. Anderson, Chemero, 2013, s. 204–205; por. podrozdział *Predykcja*). Generowane przez model predykcje pozwalają (1) kształtować i zmieniać wewnętrzny model świata, który ma rekapitulować jego strukturę przyczynową (por. Clark, 2013b, s. 182; Gładziejewski, 2016, s. 562); oraz (2) kierować odpowiednimi działaniami w ramach tzw. wnioskowania aktywnego, tj. takiego wnioskowania, jakiego wnioskiem jest określone działanie lub ich ciąg (por. podrozdziały *Wnioskowanie aktywne – Podejmowanie decyzji*). Model generatywny minimalizujący błędy predykcyjne jest uzbrojony w mechanizm predykcyjny, który realizuje określone funkcje predykcyjne. Efektem ich działania są wygenerowane predykcje, które kierują działaniami lub modyfikują wewnętrzne parametry modelu generatywnego. Rozważmy prosty przykład, który pomoże zobrazować powyższe analizy: Patrzę przez okno i widzę zachmurzone niebo. Mogę również spojrzeć na ciśnieniomierz oraz termometr. Moje wcześniejsze doświadczenia ze zmienną pogodą oraz dane dostarczone mi przez zmysły (poziom zachmurzenia, wysokość ciśnienia atmosferycznego wskazana na ciśnieniomierzu, temperatura powietrza pokazana przez termometr itd.) pozwalają mi przewidzieć, że niebawem spadnie deszcz. Informacja ta jest dla mnie kluczowa, ponieważ mam na dzisiaj umówione spotkanie w sprawie pracy i muszę być na nim elegancko ubrany oraz pojawić się w umówionym miejscu, o określonej godzinie. Wiem, że

[31] Należy zwrócić uwagę na fakt, że nie postuluję tutaj istnienia jakiejś szczególnej teorii mechanizmów normatywnych dedykowanej PP, a raczej twierdzę, że można wskazać taką klasę mechanizmów, które są normatywne, zaś podklasą tych mechanizmów są m.in. mechanizmy predykcyjne i np. mechanizmy reprezentacyjne.

załamanie pogody może utrudnić moje dotarcie na spotkanie oraz dodatkowo zniszczyć ubiór. Zatem trafność mojego przewidywania jest kluczowa dla działań, które zamierzam podjąć. Moje przewidywanie (predykcja) będące wypadkową posiadanej wiedzy oraz informacji, które mogę otrzymać ze środowiska, pozwala mi podjąć odpowiednie działania. Po pierwsze mogę zrewidować moje przekonania dotyczące pogody i uznać informacje, które otrzymałem za na tyle niepewne, że je zignoruję, w efekcie czego mogę zmoknąć. Mogę również potraktować te przewidywania na serio i podjąć takie działania, jakie maksymalnie zminimalizują skutki przewidzianego przeze mnie stanu rzeczy, czyli np. zamówić taksówkę albo zabrać parasol. Przykład ten dobrze pokazuje, że wygenerowana predykcja wpływa istotnie na wybór działań, tzn. że tworząc taką, a nie inną predykcję, jestem zobowiązany podjąć takie, a nie inne działania (Friston i in., 2010, s. 233). W tym sensie predykcja jest normatywna dla tych działań.

Wstępnie rzecz ujmując, można powiedzieć, że przez normatywność predykcji rozumiem taką ich własność, która jest konstytutywna dla kontroli i wyboru działań, a także dla struktury i treści wewnętrznego modelu świata (Bickhard, 2003; Piekarski, 2019, s. 1). Jako taka ma również wartość logiczną, którą należy wiązać z bayesowskim rozkładem prawdopodobieństwa. Może być zatem mniej lub bardziej prawdopodobna. Ta cecha predykcji czyni z nich dobre przewodniki działań, bowiem faworyzuje ona działania w tych sytuacjach, które zwiększają prawdopodobieństwo określonych predykcji, odrzucając tym samym działania, które dotyczą sytuacji przewidywanych przez predykcje o niskim prawdopodobieństwie. Predykcja jest zatem normatywna wtedy, gdy możemy jej przypisać określone wartości logiczne. Ta jej własność decyduje o tym, że może ona wpływać na treść i strukturę modelu generatywnego oraz kierować działaniami.

Nawiązując do wcześniejszych analiz, trzeba powiedzieć, że normatywność predykcji w odniesieniu do modelu generatywnego można uzasadnić internalistycznie. Inaczej jest, o czym wspominałem, omawiając funkcje reprezentowania, w przypadku normatywności w odniesieniu do działań i środowiska. Ten rodzaj normatywności, za czym będę argumentował, musi być uzasadniony eksternalistycznie. Po pierwsze należy uzasadnić, w jaki sposób przypisuje się określonej predykcji wartość logiczną[32]; oraz po drugie trzeba wskazać, w jakie relacje wchodzi ona z działaniami, które mają być realizowane w określonym środowisku. Wspomniałem wcześniej, że określona predykcja wpływa istotnie na wybór pewnych działań, faworyzując część z nich i odrzucając inne. Wróćmy na chwilę do przykładu z pogodą. Jeżeli przewiduję, że będzie padać, a chcę zminimalizować niepewność dotyczącą tego przyszłego stanu rzeczy, to w tej sytuacji optymalnym działaniem będzie zamówienie taksówki, przełożenie spotkania lub ewentualnie

[32] Można to uczynić, np. przyjmując wyjaśnienia budowane na podstawie reprezentacji strukturalnych (por. Gładziejewski, Miłkowski, 2017; podrozdział *S-reprezentacje i realne wzorce*).

zabranie parasola. Wydaje się jednak, że nie zminimalizuję zaskoczenia, jeżeli przykładowo pójdę spać, włączę telewizor bądź poczytam książkę. Działania te nie będą optymalne, ponieważ nie zwiększą mojej szansy na realizację zadania, jakie przed sobą postawiłem, czyli dotarcie na spotkanie w sprawie pracy. Można zatem powiedzieć, że predykcja normuje wybór możliwych do podjęcia działań. Schematycznie możemy to „normowanie" zapisać pod postacią następującego okresu warunkowego:

Jeżeli '*predykcja*' (warunek), to '*działanie*' (skutek).

Zależność pomiędzy predykcjami a działaniami nie jest jednak typową relacją przyczynową, którą można zapisać symbolicznie jako „Jeżeli *A*, to *B*", ale czymś, co będę określał jako „relację motywacyjną", którą należy zapisać w takiej postaci:

Jeżeli *A*, to *B, C* lub *D* (itd.), ale nie *E*, *F* lub *G* (itd.).

Predykcje bowiem warunkują pojawienie się określonych działań, wykluczając tym samym inne. Teza ta znajduje uzasadnienie w jednym z głównych założeń PP, zgodnie z którym celem funkcjonowania modelu jest minimalizacja błędów predykcyjnych. Oznacza to, że faworyzowane są tylko te działania, które sprzyjają realizacji tego celu i przyczyniają się do tego[33]. W praktyce oznacza to, że dany organizm działa w ten sposób, który zwiększa jego szansę na przeżycie, czyli zmniejsza stopień zaskoczenia powiązany z koniecznością działania w nieznanym środowisku.

Takie ujęcie jest jednak narażone na zarzut pomieszania przyczynowości z normatywnością. Można bowiem twierdzić, że opisana przeze mnie relacja motywacyjna ma charakter normatywny jedynie z punktu widzenia zewnętrznego obserwatora. Aby uniknąć tego zarzutu, należy wykazać, że relacja normatywna pomiędzy predykcjami a działaniami nie ma charakteru diadycznego, ale triadyczny, bowiem jej trzecim członem są określone stany środowiska. O konieczności uwzględnienia czynnika środowiskowego wspominał, jak już wiemy, Bickhard w swojej teorii reprezentacji jako antycypacji. Uzasadnienie normatywnego charakteru predykcji zakłada konieczność odwołania się do interakcji pomiędzy organizmem a środowiskiem, które jest, jak twierdzi wielu badaczy (por. Bruineberg, Rietveld, 2014; Bruineberg, 2017; Gibson, 1979; Norman, 2018; Piekarski, Wachowski, 2018), już wstępnie ustrukturyzowane. Należy zatem powiedzieć, że normatywność predykcji determinowana jest przez określone role funkcjonalne w modelu generatywnym (m.in. wybór i kierowanie działaniami), a także przez odniesienie do określo-

[33] Czasem jednak organizmy zachowują się w inny sposób, np. zwiększając poziom niepewności poprzez wybór takich działań, jakie są ryzykowne. Por. podrozdział *Problem ciemnego pokoju*.

nych właściwości środowiska naturalnego, ale również społeczno-kulturowego. Dzięki temu, że świat jest już „uprzednio" ustrukturyzowany, może przedstawiać organizmowi określone wartości nagrody i kary, których, co wykażę w rozdziale *Mechanizmy normatywne a działania w przetwarzaniu predykcyjnym*, nie da się zredukować do negatywnego zaskoczenia (*negative suprise*) lub zwiększenia wierzytelności modelu (*Bayesian log-evidence*) (por. Friston, Adams, Montague, 2012; por. także podrozdział *Podejmowanie decyzji*).

Innymi słowy, predykcje są normatywne, ponieważ odnoszą się nie tylko do konieczności minimalizowania błędów predykcyjnych i niepewności, ale także do indywidualnych przekonań, pragnień lub motywacji, które powstają w odniesieniu do określonych stanów środowiska, np. afordancji (por. Gibson, 1979). Z tej perspektywy ewentualny błąd lub błędna reprezentacja ma normatywne znaczenie dla organizmu, ale nie tylko jako potencjalny skutek określonych procesów przyczynowych. Normatywny charakter predykcji polega przede wszystkim na tym, że w istotny sposób kształtują one przyczynowe przejścia pomiędzy stanami, które mają określoną treść, a ustrukturyzowanym środowiskiem, determinując tym samym wybór odpowiednich działań w taki sposób, że odpowiadają one normatywnej regule Bayesa (por. Hohwy, Roepstorff, Friston, 2008; Kiefer, 2017; Shams, Ma, Beierholm, 2005). Zatem to, co dany organizm robi, zależy nie tylko od konieczności minimalizacji błędów predykcyjnych, indywidualnych preferencji, potrzeb i przekonań, ale także od określonych własności środowiska. Oznacza to, że działania są wybierane na podstawie pewnego warunkowego potencjału (powiązanego z predykcjami) oraz relacji, w które dany organizm wchodzi ze środowiskiem. Należy zatem zgodzić się z Bickhardem, który twierdzi, że „takie warunkowe relacje mogą rozgałęziać się – pojedynczy wynik interakcji może funkcjonować w celu wskazania wielu dalszych interakcyjnych możliwości – i mogą one iterować – ukończenie interakcji *A* może wskazywać na możliwość *B*, która, jeśli zostanie ukończona, wskazywałaby na możliwość *C* i tak dalej" (Bickhard, 2009, s. 78). Znaczy to, że system poznawczy nadal przewiduje formę i treść sygnałów zmysłowych, nie tylko wykorzystując posiadaną wiedzę i reguły wnioskowania, ale także na drodze aktywnego działania w środowisku.

Pełnemu uzasadnieniu powyższych tez jest poświęcony rozdział *Mechanizmy normatywne a działania w przetwarzaniu predykcyjnym*. Teraz powrócę do problemu wyjaśnień formułowanych na gruncie PP oraz pytania dotyczącego normatywności bayesowskiego modelu rozwijanego przez badaczy pracujących w tym nurcie (por. podrozdział *Przetwarzanie predykcyjne jako bayesizm oświecony*).

4.5. Przetwarzanie predykcyjne a mechanistyczny model wyjaśniania naukowego

W rozdziale *Przetwarzanie predykcyjne jako bayesowski model eksplanacyjny* stwierdziłem, że PP, w przeciwieństwie do innych modeli typu bayesowskiego, np. *analiza racjonalna*, w nietrywialny sposób pretenduje do bycia teorią wyjaśniającą poznanie, która formułuje opisy na każdym z trzech poziomów analizy wyróżnionych przez Marra. Z jednej bowiem strony wielu badaczy (por. m.in. Gładziejewski, 2019; Hohwy, 2015a) podaje szereg argumentów za mechanistycznym odczytaniem PP, z drugiej zaś strony, co już uzasadniałem, PP jest takiego rodzaju modelowaniem bayesowskim, które ma charakter normatywny w sensie epistemicznym. Oznacza to, że PP oferuje możliwość wyjaśniania określonych zjawisk jako aproksymacji wnioskowania bayesowskiego. W ujęciu tzw. preskryptywizmu epistemicznego nie jest przesądzone, czy zjawiska wyjaśniane w ramach PP są bayesowskie z natury. Pełne uzasadnienie realizmu w odniesieniu do bayesowskiego PP zakłada przyjęcie tzw. mocnego preskryptywizmu (ontycznego), który zakłada, że zjawiska wyjaśnia się przy pomocy twierdzenia Bayesa, ponieważ są one bayesowskie z natury. Jeżeli są one bayesowskie z natury, to znaczy, że *powinno* się je wyjaśniać przy użyciu modelowania Bayesa. W niniejszych rozważaniach powracam do pytania o to, czy preskryptywizm ontyczny jest uzasadniony w odniesieniu do PP, a jeżeli jest, to czy mechanistyczna teoria wyjaśnień oferuje adekwatny i nietrywialny model wyjaśnień dla tej ramy badawczej. Aby odpowiedzieć na powyższe pytania, wpierw uzasadnię, w jaki sposób przyjęta w tym rozdziale koncepcja normatywności mechanizmów predykcyjnych pozwala uzasadnić realizm w odniesieniu do bayesowskiego PP, a następnie ogólnie omówię mechanistyczny model wyjaśnień w kognitywistyce i będę bronił tezy, że jeżeli PP wyjaśnia (w jakimś stopniu) zjawiska poznawcze, to wyjaśnia jej poprzez opis mechanizmów percepcyjnych i poznawczych.

4.5.1. Normatywność mechanizmów predykcyjnych a preskryptywizm ontyczny

Zgodnie z bronioną przeze mnie tezą preskryptywizmu ontycznego, w odniesieniu do bayesowskiego PP, wyjaśnianie zjawisk poznawczych przy użyciu pojęć modelu generatywnego, mechanizmów i funkcji predykcyjnych, inferencji itd. ma charakter ontyczny, tzn. istnieją mocne przesłanki przemawiające za tym, że wyjaśnienia formułowane na gruncie PP są bayesowskie, ponieważ bayesowskie z natury są niektóre funkcje poznawcze. Mówiąc, że dany mechanizm czy proces jest aproksymacją wnioskowania bayesowskiego, głosimy tak, nie z powodów pragmatycznych czy instrumentalnych, ale dlatego, że te mechanizmy oraz procesy są

bayesowskie, ponieważ takie są. Wsparciem dla tej tezy są przeprowadzone tutaj badania nad normatywnością funkcji i mechanizmów. Mianowicie: pokazałem, że nietrywialne wyjaśnienie sposobu, w jaki percepcja, oparta na bayesowskich mechanizmach predykcyjnych i strukturze modelu generatywnego, wiąże się z działaniami podejmowanymi przez dany organizm w środowisku, wymaga odwołania się do normatywności „wpisanej" w te mechanizmy. Normatywność ta z jednej strony wiąże się ze strukturą i wewnętrznymi parametrami modelu generatywnego wykorzystującego wnioskowanie probabilistyczne typu bayesowskiego w celu wygenerowania wysoce prawdopodobnych predykcji kierujących działaniami, z drugiej zaś ze strukturą i własnościami środowiska, w którym organizm żyje i funkcjonuje. Odrzucenie tezy o normatywności mechanizmów predykcyjnych i predykcji prowadzi do przyjęcia prostych wyjaśnień przyczynowych, które z jednej strony w satysfakcjonujący sposób pokazują, jak działa model generatywny w ujęciu etiologicznym i w odniesieniu do określonego typu (np. gatunku), z drugiej zaś nie dają jednak odpowiedzi na pytania dotyczące funkcjonowania i działania określonego indywiduum w konkretnej sytuacji środowiskowej. Jeżeli bowiem odrzuci się bronioną tutaj tezę o normatywności niektórych procesów percepcyjnych, to (1) życie i działanie organizmów sprowadza się tylko do wymogu minimalizacji błędów predykcyjnych oraz niepewności; (2) neguje się możliwość wyjaśnienia takich działań, które zwiększają stopień niepewności zamiast go redukować (problem ciemnego pokoju czy kompromis eksploracji i eksploatacji por. podrozdział *Problem ciemnego pokoju*); (3) problematyczne staje się wyjaśnienie błędów poznawczych i wartości logicznej predykcji oraz reprezentacji, oraz tego, dlaczego jest ona istotna dla działającego w środowisku organizmu (por. Bickhard, 2003, 2016).

Teza o normatywności, która stanowi dla mnie mocną przesłankę za przyjęciem stanowiska realistycznego w odniesieniu do PP, jest ugruntowana w przekonaniu, zgodnie z którym mechanizmy są normatywne, o ile pozwalają one wyjaśnić normatywność określonych funkcji (por. Garson, 2013)[34]. W tym sensie twierdzę, że można o nich mówić jako o mechanizmach normatywnych. Garson podkreśla, że jeżeli nie odwołamy się do normatywnych funkcji realizowanych przez mechanizmy, to trudne staje się wyjaśnienie ich dysfunkcji, co w efekcie może prowadzić do mówienia o mechanizmach odpowiedzialnych za dysfunkcje, np. mechanizmach, które odpowiadają za zawały serca, niedziałanie mikserów czy błędne reprezentacje. Teza głoszona przez Garsona sugeruje, że normatywność nie jest tylko pojęciem użytecznym pragmatycznie albo ważnym z powodu określonej strategii badawczej,

[34] Można postawić bronionemu tutaj ujęciu mechanizmów normatywnych zarzut, że problem normatywności mechanizmów sprowadza się do problemu funkcji normatywnych, zatem wystarczy, że będzie się mówiło o mechanizmach funkcjonalnych. Twierdzę jednak, że bronione przeze mnie pojęcie mechanizmów normatywnych jest zasadne, bowiem każdy mechanizm normatywny jest mechanizmem funkcjonalnym, ale nie każdy mechanizm funkcjonalny jest mechanizmem normatywnym. Dlaczego? Ponieważ mechanizm normatywny w moim sensie wiąże się ze wskazaniem przyczyn konstytutywnych dla danego zjawiska, a nie tylko ze wskazaniem określonych funkcji. Twierdzenie to w dalszych rozważaniach zostanie w pełni wyeksplikowane.

ale jest pojęciem, które wydaje się, że ma określoną moc eksplanacyjną[35]. Niemniej można Garsonowi postawić zarzut podobny do tego, który stawia się koncepcjom teleosemantycznym: pojęcia funkcji i dysfunkcji konstytutywnie się zakładają i współdefiniują, nie ma bowiem funkcji bez dysfunkcji (por. podrozdział *Funkcja reprezentowania*). Z tego też powodu trudno mówić o tym, że takie pojęcie funkcji coś faktycznie wyjaśnia.

Można jednak twierdzić, i tego ujęcia będę bronił, że jeżeli orzeka się, iż dany mechanizm albo funkcja jest normatywny, to ma się na myśli, że odgrywa on określone role przyczynowe. Twierdzę zatem, że mówiąc o takich rolach w odniesieniu do normatywnych własności mechanizmów i funkcji predykcyjnych, mówi się o nich jako przyczynach określonych działań danego organizmu w środowisku. Innymi słowy, normatywność jest takim predykatem, za pomocą którego możemy wyjaśnić pewne zjawiska (tj. działania i zachowania) w kategoriach mechanizmów i funkcji, które są ich przyczynami.

Podsumowując: jeżeli wyjaśnienia formułowane na gruncie PP mają charakter normatywny (a część z nich ma, co już pokazałem), to są one takie z powodu istnienia określonych mechanizmów i funkcji normatywnych. Zatem wyjaśnienia te są przedmiotowo ważne i nietrywialne nie tylko z powodów epistemicznych, ale także ontycznych. Oznacza to, że zasadne jest przyjęcie stanowiska preskryptywizmu ontycznego w odniesieniu do PP, czyli przekonania o realistycznym charakterze wyjaśnień formułowanych na gruncie tej ramy badawczej.

Po tych ustaleniach możemy zapytać o sposób, w jaki PP wyjaśnia zjawiska poznawcze. Będę argumentował za tym, że jest to praktyka badawcza, którą można ująć w kategoriach odkrywania i opisywania mechanizmów odpowiedzialnych za realizację określonych funkcji poznawczych. Zanim to jednak uczynię, konieczna jest ogólna charakterystyka teorii wyjaśnień mechanistycznych.

4.5.2. Mechanistyczny model wyjaśniania naukowego a mechanizmy normatywne

Praktyki badawcze w nauce można opisać w kategoriach odkrywania i opisywania mechanizmów. Aby wyjaśnić dane zjawisko, należy poszukać mechanizmu, który odpowiada za jego realizację. Wyjaśnienia mechanistyczne są stosowane z dużymi sukcesami w naukach biologicznych, neuronaukach, ale także w naukach społecznych (por. Glennan, Illari, 2018, część IV)[36]. Opierają się one na dwóch podstawowych założeniach: (1) należy odróżnić zjawisko, które ma być wyjaśnione (*explanan-*

[35] Na temat pojęcia mocy eksplanacyjnej por. Kaplan, Craver, 2011.

[36] W niniejszej koncepcji wyjaśnień mechanistycznych korzystam przede wszystkim z ujęcia proponowanego przez Cravera. Należy również dodać, że współczesny mechanicyzm określa się mianem „nowego mechanicyzmu" lub po prostu „neomechanicyzmu".

dum), od konkretnego systemu, który leży u podstaw tego zjawiska (*explanans*)[37]; (2) mechanizm, przy którego pomocy dane zjawisko ma zostać wyjaśnione, jest wykrywalny i opisywalny. Kluczowe dla wyjaśnień mechanistycznych jest opisane przez Kaplana ograniczenie 3M (model – mechanizm – mapowanie). Zakłada ono, że model zjawiska docelowego wyjaśnia to zjawisko wtedy, gdy (1) zmienne w modelu odpowiadają możliwym do zidentyfikowania komponentom, działaniom oraz własnościom organizacyjnym mechanizmu docelowego, który wytwarza, utrzymuje lub leży u podstaw tego zjawiska; oraz (2) zależności (być może matematyczne) znajdujące się wśród tych (być może matematycznych) zmiennych w modelu odpowiadają relacjom przyczynowym pomiędzy komponentami mechanizmu docelowego (Kaplan, 2011, s. 347). Oznacza to, że zjawisko z jednej strony składa się z pewnych funkcjonalnych komponentów[38], którym można przypisać odpowiednie operacje, zaś z drugiej mechanizm, który wyjaśnia dane zjawisko, jest rozkładalny na poszczególne zintegrowane ze sobą części, które odpowiadają za dane zjawisko (Zednik, 2008, s. 1454). W takim ujęciu przez mechanizm należy rozumieć „strukturę, która pełni daną funkcję dzięki swoim komponentom, ich operacjom oraz organizacji. Takie skoordynowane funkcjonowanie mechanizmu jest odpowiedzialne za powstanie jednego lub więcej różnych zjawisk" (Bechtel, Abrahamsen, 2005, s. 423; Bechtel, 2008, s. 13). Carl F. Craver (2007, s. 6) proponuje, aby mechanizmy rozumieć jako „jednostki i działania zorganizowane w taki sposób, że razem egzemplifikują one zjawisko, które stanowi eksplanandum"; zaś Peter Machamer, Lindley Darden i Carl F. Craver (2011, s. 147) twierdzą, że „mechanizmy są to przedmioty i działania zorganizowane w taki sposób, że wywołują regularne zmiany od początku, albo stanu wyjściowego, do końca, albo warunków zakończenia". Ogólnie mówiąc, należy stwierdzić, że mechanizmy składają się z przedmiotów (komponenty), ich działań i czynności (operacje) oraz relacji pomiędzy nimi. Tworzą one wielopoziomowe hierarchie, które wiążą się z ich organizacją. Przedmioty są tymi elementami mechanizmów, które uczestniczą w działaniach. Warunkiem ich działania są określone typy własności danych przedmiotów. Operacja realizowana przez komponent określonego mechanizmu może stanowić eksplanandum i zostać wyjaśniona przez mechanizm znajdujący się na niższym poziomie hierarchii. Przykładowo działanie, jakim jest pompowanie krwi przez serce w ramach mechanizmu odpowiedzialnego za dystrybucję tlenu do komórek organizmu, można potraktować jako eksplanandum, które da się wyjaśnić mechanicznie. Takie działanie można wyjaśnić np. poprzez odwołanie się do mechanizmu z niższego poziomu organizacji, jakim jest zorganizowany układ skurczających i rozkurczających się komór serca (Gładziejewski, 2015, s. 74).

[37] Należy stwierdzić, że jest to założenie wspólne dla wielu modeli wyjaśniania, począwszy od modelu Carla G. Hempla i Paula Oppenheima.

[38] Mogą oczywiście istnieć złożone mechanizmy, które nie składają się z funkcjonalnych komponentów, np. punkt geometryczny bez wyróżnionych części.

Machamer, Darden i Craver twierdzą, że mówienie o przedmiotach i ich działaniach nie jest najbardziej adekwatne dla badania oraz opisywania mechanizmów. Zamiast tego proponują, aby korzystać z pojęcia funkcji. Funkcje zaś to nic innego niż role, które są odgrywane przez przedmioty i działania w mechanizmie. Rozpoznać funkcję to tyle, co zauważyć, że jest ona składnikiem określonego mechanizmu. Przez funkcję należy raczej rozumieć działania, za pośrednictwem których przedmioty przyczyniają się do pracy mechanizmu, aniżeli jakieś własności (Machamer, Darden, Craver, 2011, s. 151). Jeżeli funkcje są działaniami, to w jaki sposób należy rozumieć działania? Działania to rodzaje przyczyn. Przedmiot może pełnić funkcję przyczyny, jeżeli uczestniczy w jakimś produktywnym działaniu. W praktyce oznacza to, że to nie przedmioty czy rodzaje naturalne są przyczynami w sensie pierwotnym, tylko to, co one robią, czyli ich działania. Działaniem serca jest pompowanie krwi i w tym sensie serce jest przyczyną dotleniania mózgu. Mówiąc o przyczynie w sensie pierwotnym, ma się na myśli jego funkcję, jaką jest pompowanie krwi (w przypadku serca).

Craver (2012) podaje mocny argument na rzecz powiązania wyjaśnienia mechanizmów z ich funkcjami (por. także Piccinini, Craver, 2011). Zwraca on uwagę na to, że pomimo odrzucenia przez wiele nauk wyjaśnień teleologicznych, to zarówno nauki fizjologiczne, jak i neuronauki, często powołują się na opisy funkcjonalne. One bowiem niejednokrotnie prowadzą do poszukiwania mechanizmów. Opisy funkcjonalne przyczyniają się do wyjaśnień mechanistycznych na trzy sposoby: (1) jako sposoby starannego wskazywania odpowiednich wyjaśnień etiologicznych; (2) jako sposoby ustanawiania (*framing*) wyjaśnień konstytutywnych[39]; oraz (3) jako sposoby wyjaśniania określonych przedmiotów przez umiejscawianie ich w mechanizmach wyższego rzędu. Istotne jest to, że opisy funkcjonalne mają charakter ontyczny, co oznacza, że nie opierają się one na decyzjach i strategii badawczej obserwatora, tylko na faktycznych regularnościach obecnych w zjawiskach (por. także Craver, 2013).

4.5.2.1. Mechanizmy predykcyjne jako przyczyny a subosobowy poziom wyjaśniania

Docieramy tutaj do wniosku naszych rozważań, według którego mechanizmy realizujące funkcje normatywne mogą być przyczynami określonych działań i zachowań. Rozpoznajemy je na podstawie identyfikacji takich, a nie innych funkcji. Istotne jest to, że nie wszystkie mechanizmy mogą być konstytutywnymi

[39] Wyjaśnienia konstytutywne polegają na tym, że wyjaśnia się dane zjawisko, wskazując jego strukturalną (wewnętrzną) przyczynę. W tym sensie takie wyjaśnienia przeciwstawia się wyjaśnieniom etiologicznym (Craver, 2013, s. 151).

przyczynami działań i zachowań, ale tylko te, które można określić normatywnymi. Oznacza to, że odróżniam mechanizmy normatywne, czyli konstytutywne przyczynowo, od mechanizmów po prostu przyczynowych. Z tego też powodu nie można stwierdzić, że mechanizm odpowiedzialny za poruszanie ręką jest faktyczną, tj. konstytutywną przyczyną tego, że np. chciałem się napić kawy z filiżanki. Jest to oczywiście mechanizm, który współkonstytuuje czynność złapania filiżanki i skierowania jej w stronę ust, niemniej dużym nadużyciem byłoby stwierdzenie, że jest on przyczyną całego działania, które można określić mianem picia kawy. Inaczej jest w przypadku (normatywnego) mechanizmu predykcyjnego (który określa przyczynę konstytutywną danego działania): poczucie pragnienia i zmęczenia zwiększa niepewność w środowisku. Jeżeli jestem senny i spragniony, to moje możliwości działania są dużo mniejsze, co w efekcie może wpłynąć na pojawienie się sytuacji o większym stopniu niepewności. Predykcja, zgodnie z którą zminimalizuję potencjalną niepewność, jeżeli wypiję napój zawierający kofeinę, może prowadzić do działania, jakim jest spożywanie kawy. Pójdźmy dalej. Odpowiedni wysokopoziomowy mechanizm predykcyjny odpowiedzialny za pojawienie się predykcji minimalizujących błędy predykcyjne, powiązane z niską wydajnością organizmu, stanowi przyczynę takiego, a nie innego działania, w tym przypadku picia kawy. Odróżniam jednak bycie przyczyną danego działania od bycia jego racją. Pomieszanie bowiem tych pojęć może skutkować fałszywym potraktowaniem PP w kategoriach zdroworozsądkowej psychologii potocznej (*folk psychology*)[40]. Aby uzasadnić to przekonanie, odwołam się do dwóch klasycznych propozycji analizujących pojęcia racji i przyczyny, tj. koncepcji Donalda Davidsona i Elizabeth Anscombe. Rozpocznę od autora pierwszego.

Zdaniem Davidsona (1963) racjonalizacja jest jedną z odmian wyjaśnienia przyczynowego. Jeżeli ktoś zrobił coś z powodu pewnych racji, to można go scharakteryzować jako (1) mającego pewne przednastawienie (*pro attitude*) lub pewną postawę do działania (w taki, a nie inny sposób)[41]; oraz (2) mającego dane przekonanie (wiedzącego coś, spostrzegającego coś czy pamiętającego coś itd.). Podanie racji dla działania, które działający wykonał, jest zatem wskazaniem pewnego przednastawienia i przekonania podzielanych przez działającego. Parę nastawienie/przekonanie Davidson określił mianem „racji pierwotnej" bądź „podstawowej" (*primary reason*). Racja podstawowa wyjaśnia, dlaczego działający podjął się danego działania. Podanie jej jest konieczne i wystarczające dla zrozumienia tego, w jaki sposób racja dowolnego rodzaju racjonalizuje działanie. Davidson stwierdza, że

[40] Psychologię potoczną w najprostszy sposób można scharakteryzować jako: (PP1) zdroworozsądkowy sposób rozumienia różnych zjawisk psychicznych, a także pewien sposób przypisywania innym przekonań, pragnień, intencji czy emocji; oraz jako (PP2) podstawowy sposób wyjaśniania i przewidywania ludzkich działań (Ravenscroft, 2019).

[41] Przednastawienia Davidson rozumie jako pragnienia, chcenia, popędy, podszepty, moralne i estetyczne przekonania, ekonomiczne przedsądy, konwencje społeczne, publiczne i prywatne cele itd.

z tego powodu należy ją traktować jako przyczynę danego działania. Rozważmy prosty przykład: naciskając włącznik, zapalam światło i pokój zostaje oświetlony. Nacisnąłem włącznik, ponieważ chciałem zapalić światło. Mówiąc, że chciałem zapalić światło, podaję rację, czyli racjonalizuję to, że nacisnąłem włącznik. Podając rację, nie racjonalizuję jednak zapalenia światła albo oświetlenia pokoju. Racje mogą racjonalizować to, co ktoś robi wtedy, kiedy jest to opisane w taki sposób, że podane zostają konieczne warunki dla podstawowej racji. Na podstawie analizy tego przykładu Davidson stwierdza: „*R* jest racją podstawową tego, że działający wykonał dane działanie *A* zgodne z opisem (*under the description*) *d*, wtedy i tylko wtedy, gdy *R* składa się z nastawienia działającego do odpowiedniego działania z pewną własnością oraz jego przekonania, że *A* pod opisem *d* zawiera tę własność" (Davidson, 1963, s. 687). Dla Davidsona przyczyną działania jest jego racja podstawowa, którą można opisać w kategoriach przekonań i nastawień, co w praktyce oznacza, że przyczyną (fizycznego) działania jest jakiś niefizyczny stan będący ucieleśnieniem pary pragnienie/przekonanie. Są to pojęcia tzw. psychologii potocznej.

Anscombe (1957) zaproponowała inne ujęcie, zupełnie zrywające z rozwiązaniem Davidsona. Jej zdaniem odróżnienie działań od niedziałań ma sens tylko wtedy, gdy odpowiedź na pytanie „dlaczego?" wskazuje na jakiś motyw: przeszłe wydarzenie, interpretacje konkretnego działania czy jakieś wydarzenie przyszłe. Nie mają sensu odpowiedzi, w których padają stwierdzenia w stylu: „Nie byłem świadomy, gdy to robiłem", „Zaobserwowałem to, co robiłem" albo kiedy odpowiedź zawiera świadectwo lub wskazanie jakiejś przyczyny, także mentalnej. Musimy – i to jest kluczowy argument Anscombe – umieć odpowiedzieć na pytanie „dlaczego?", odnosząc się do tego, co było zamierzone (*intend*), a nie do tego, co zostało intencjonalnie dokonane. Odpowiedź na pytanie „dlaczego?" powinna być sprowadzalna do formy: „Jeżeli ja robię *p* z intencją *Q*, to *Q*". Zwrot „z intencją" wskazuje na wewnętrzną teleologiczną strukturę działania.

Anscombe przytacza sławny przykład człowieka, który pompuje wodę do domu, poruszając ręką raz w górę, raz w dół (Anscombe, 1957, s. 37–38). Woda jest zatruta, a obsługujący pompę nie wie o tym. Jest przekonany, że dostarcza czystą wodę mieszkańcom domu. Możemy go zapytać: „Dlaczego wykonujesz działanie *X*?". W tym przypadku pytanie będzie brzmiało: „Dlaczego poruszasz swoją ręką z góry do dołu?". Na tak zadane pytanie mężczyzna może podać następujące odpowiedzi: (1) „Pompuję"; (2) „Pobieram wodę ze studni"; (3) „Dostarczam wodę mieszkańcom domu". Nie może jednak stwierdzić, że (4) „Chcę otruć mieszkańców domu". Dla Anscombe szczególnie interesujące jest zdanie (3), ponieważ skierowane jest ono nie na to, co mężczyzna robi, ale na to, co chce czy zamierza zrobić lub osiągnąć. Wyraża ono intencję skierowaną na przyszłość, intencję działania, a nie intencję, z którą wykonuje dane działanie. Anscombe powiada: w przypadku

mężczyzny pompującego wodę mamy do czynienia z jednym działaniem, które można opisać na kilka sposobów. Każdy z nich zależy od różnych okoliczności i każdy jest opisem umożliwiającym opis następujący po nim. Ludzkie działania, konkluduje Anscombe, mają podobną do siebie strukturę teleologiczną (czasem prostą, czasem bardziej złożoną), która składa się z różnych faz tego, co dana osoba robi. W tym przykładzie działanie będzie intencjonalne pod opisem (*under description*) „Dostarczam wodę mieszkańcom domu", a nie będzie intencjonalne pod opisem „Chcę otruć mieszkańców domu". Człowiek intencjonalnie dostarcza wodę mieszkańcom domu, a nieintencjonalnie ich zatruwa[42].

W proponowanym przez filozofkę ujęciu struktura działania intencjonalnego jest jednocześnie strukturą wyjaśniającą je. Odpowiedź na pytanie o to, dlaczego dana osoba robi to a to, odnosi się wprost do wewnętrznej racji wyjaśniającej działanie. Intencja pompującego nie pełni tutaj funkcji określającej jakiś wewnętrzny stan jego umysłu, ale funkcję przyczynowo-skutecznego działania. Intencja jako predykcja jest zatem racją danego działania. Koncepcja Anscombe wiąże racje z predykcyjnym wymiarem intencji praktycznej.

Proponowane przeze mnie ujęcie również zrywa z koncepcjami Davidsona i Anscombe. Po pierwsze racja jest przez nich w pewnym sensie utożsamiana z intencją. Intencja jednak nie ma charakteru mentalnego ani poznawczego, ale ujmowana jest przede wszystkim w zdaniach wyrażających zamiary działającego, czyli istotny jest jej językowy wyraz. Działanie zaś jest intencjonalne pod pewnym opisem. Wydaje się, że takie rozwiązanie w przypadku Anscombe pozwala uniknąć sprowadzania wyjaśnienia działań do języka psychologii potocznej, niemniej, kiedy przeanalizuje się podane przez nią przykłady, to można odnieść wrażenie, że wskazanie intencji jest ostatecznie, w jakimś stopniu, zależne od stanu wiedzy działającego: mężczyzna wiedział, że pompuje wodę, ale nie wiedział, że zatruwa mieszkańców[43].

Ujęcia Anscombe i Davidsona oferują pewien sposób wyjaśniania zjawisk. Wiąże się on z podaniem racji czy motywu, który pozwala nam mówić o tym, że dane działanie jest normatywne bądź racjonalne. Są to jednak wyjaśnienia obce PP. Głównym powodem jest to, że broniona przeze mnie koncepcja zrywa z potoczno-psychologicznym myśleniem o zjawiskach poznawczych[44]. Rama bayesowska PP

[42] W pewnym sensie Anscombe antycypuje davidsonowską tezę mówiącą, że wszystkie działania są intencjonalne pod pewnym opisem. Dane działanie może być intencjonalne pod jednym opisem, pod innym zaś nie.

[43] Niebanalne znaczenie ma tutaj stwierdzenie, że wiedza o intencjach nie ma charakteru obserwacyjnego. Jest ona raczej *wiedzą-jak*, a nie operacyjną *wiedzą-że*. Z tego powodu nie można jednoznacznie zakwalifikować Anscombe do obozu myślicieli uznających ważną rolę eksplanacyjną pojęć z zakresu psychologii potocznej. Ten wątek musi jednak pozostać na uboczu niniejszych rozważań.

[44] Inaczej twierdzi Joe Dewhurst (2017), utrzymując, że PP zrywa jedynie z dosłowną interpretacją psychologii potocznej jako teorii postaw propozycjonalnych, jest jednak zgodne z jej rozumieniem jako „szerszej" perspektywy badania zjawisk poznawczych. Ze względu na ograniczenia nałożone na nasze rozważania, nie będę w tym miejscu dyskutował poglądu tego autora.

wiąże bowiem przekonania z predykcjami o określonej wartości w odniesieniu do rozkładu prawdopodobieństwa oraz dotyczy, jak dotąd, subosobowego poziomu wyjaśniania ludzkiego poznania. Psychologię potoczną zaś należy wiązać z poziomem osobowym[45]. Odróżnienie to zaproponował Dennett. Jego (1995b, s. 106–107) zdaniem:

> Gdybyśmy powiedzieli, że osoba posiada doznanie bólu, umiejscawia je i skłonna jest do reagowania w pewien sposób, powiedzielibyśmy 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.

Gładziejewski słusznie zauważa, że niefortunne jest to, że Dennett jako przykład stanu osobowego podaje doznanie bólu, a nie dowolną postawę propozycjonalną (Gładziejewski, 2015, s. 298). Taki przykład bowiem osłabia moc jego argumentacji. Możemy sobie wyobrazić, że w powyższym cytacie zamiast mówienia o doznaniu bólu, mówi się o przekonaniu, że będzie padać. To, co jest istotne w odróżnieniu obu tych poziomów, to stwierdzenie, że wyjaśnienia na poziomie osobowym prowadzą do odpowiedzi na pytania *dlaczego*, zaś ich eksplanandami są działania. Explanantia zaś odwołują się do stanów mentalnych. Dobrze to widać na przykładzie koncepcji Anscombe. Inaczej jest z wyjaśnieniami formułowanymi na poziomie subosobowym. Po pierwsze wyjaśniają ono to, co się dzieje, wskazując np. określony mechanizm neuronalny albo motoryczny; po drugie wskazują one określone warunki wewnętrzne, które umożliwiają posiadanie takich, a nie innych zdolności; oraz po trzecie *explanantia*, na które się tutaj powołujemy, to nie tyle stany mentalne, ile komponenty lub stany komponentów ośrodkowego układu nerwowego (Gładziejewski, 2015, s. 298–300). W odniesieniu do przeprowadzonych przeze mnie analiz należy uzupełnić ostatnie zdanie o podkreślenie znaczenia określonych stanów środowiska, w którym dany organizm funkcjonuje. Ważność tego elementu zostanie jeszcze przeze mnie mocniej uzasadniona. José L. Bermúdez (2005, s. 31–35) powiada, że wyjaśnienia osobowe są horyzontalne, czyli wewnątrzpoziomowe. Znaczy to, że *explanantia* znajdują się na tym samym poziomie, co eksplananda; czasowo je poprzedzają oraz są w jakimś sensie ich przyczynami. Wyjaśnienia subosobowe są wertykalne, czyli międzypoziomowe. Zatem zjawiska z wyższego poziomu wyjaśnia się poprzez odwołanie do zjawisk niższego poziomu.

[45] W sprawie tego odróżnienia por. również Drayson, 2014.

Ta ostatnia uwaga wiążę wyjaśnienia na poziomie subosobowym z mechanistycznym modelem wyjaśniania w nauce ze względu na hierarchiczny i wielopoziomowy sposób wyjaśniania oraz z powodu odrzucenia języka psychologii potocznej, tj. postaw propozycjonalnych oraz mówienia o stanach fizycznych i przeciwstawianych im stanach mentalnych. Uzupełniając te uwagi i dookreślając moje wcześniejsze stwierdzenie, należy powiedzieć, że jeżeli chce się wyjaśnić dane działania czy zachowania na poziomie subosobowym, to należy wskazać taki stan mentalny, przekonanie, pragnienie, intencję lub inne działanie, które stanowi jego rację. Jeżeli zaś chce się wyjaśnić dane działanie czy zachowania na poziomie subosobowym, należy wskazać taki mechanizm, proces bądź funkcję, który stanowi jego przyczynę, tzn. jest tym, co sprawia, że dane zjawisko zostaje zrealizowane[46]. W bronionej przeze mnie koncepcji przyczynami działań są określone mechanizmy predykcyjne realizujące takie, a nie inne funkcje normatywne w odniesieniu do konieczności minimalizacji błędów predykcyjnych. Nie są to, rzecz jasna, jedyne przyczyny tych działań, są to jednak takie przyczyny, jakich podanie wyjaśnia powodzenie lub porażkę działania, które podjął dany organizm w określonym środowisku i sytuacji. Są to zatem przyczyny, które określam mianem „konstytutywnych".

4.5.2.2. Możliwość mechanistycznego wyjaśnienia normatywności

Bickhard (2007) występuje przeciwko głoszonej przeze mnie tezie, zgodnie z którą funkcje normatywne (np. funkcja predykcyjna modelu generatywnego) można wyjaśnić poprzez opis i rekonstrukcję mechanizmów odpowiedzialnych za realizację takich funkcji. Jego zdaniem wskazanie samych mechanizmów jest niewystarczające dla wyjaśnienia normatywności bądź reprezentacji. Bickhard twierdzi, przypomnijmy, że reprezentacje są normatywne, ponieważ mają warunki prawdziwości. Jest to również jeden z warunków, jaki musi spełniać normatywna funkcja predykcyjna. Jeżeli zatem o normatywności reprezentacji, *resp.* funkcji, decyduje możliwość przypisania jej nietrywialnych warunków prawdziwościowych (por. Burge, 2010), to teoria, która ma wyjaśnić pojawienie się funkcji normatywnych, powinna móc zdać sprawę z czegoś, co Bickhard nazywa „asymetrycznym rozróżnieniem" (Bickhard, 2007, s. 581). Polega ono na tym, że niektóre zróżnicowane możliwości lub kategorie są lepsze od innych pod jakimś względem. Jego zdaniem w naukach takich jak fizyka pewne pojęcia czy oceny są już normatywne niejako „z góry" i dlatego ta nauka nie pozwala uwzględnić niezależnego pojawienia się zjawisk normatywnych. Powodem tego ma być liniowy charakter ujmowania

[46] Nie będę w tym miejscu rozstrzygał, jakie korzyści daje każde z tych wyjaśnień. Na boku pozostawię również kwestię relacji pomiędzy poziomami osobowym i subosobowym. Istnieją jednak pewne argumenty, zgodnie z którymi podział osobowe/subosobowe jest nieadekwatny względem PP. Por. Piekarski, 2017a.

organizmów żywych i abstrahowania od asymetrycznych różnic pomiędzy nimi a środowiskiem. Jedynym wyjątkiem jest, zdaniem Bickharda, termodynamika. Obejmuje ona bowiem rodzaje nieliniowości, które mogą leżeć u podstaw asymetrycznych różnic. Wiąże się to z faktem, że organizmy żywe opierają swoją stabilność na pozostawaniu w stanach dalekich od równowagi termodynamicznej ze środowiskiem (por. także Friston, 2010). Ta asymetria pozwala wyjaśnić pojawienie się normatywności. Najważniejsze w przypadku żywych systemów jest to, że unikają one równowagi termodynamicznej nie tylko w odniesieniu do swoich „działań", ale także w odniesieniu do swojego istnienia. Organizm, który osiągnie równowagę termodynamiczną, umiera. Pojawia się zatem taka asymetria na linii organizm – środowisko, która jest niezbędna dla pojawienia się funkcji normatywnych. Z tego też powodu, konkluduje Bickhard, ujęcie normatywności za pośrednictwem mechanizmów nie jest satysfakcjonujące.

Stanowisko Bickharda budzi pewne zastrzeżenia. Przede wszystkim nie jest jasne, co ten autor rozumie przez mechanizm. Wydaje się, że pojęcie to wiąże z mechaniką opartą na pierwszej zasadzie termodynamiki. Zasada ta głosi, że energia wewnętrzna układu może się zmieniać jedynie dzięki oddziaływaniu z otoczeniem. Oznacza to, że energia w układzie izolowanym nie może być ani utworzona, ani zniszczona, ponieważ możliwe są tylko przemiany jednych form energii w inne. Przykładem takiego stanu rzeczy jest proces przemiany energii chemicznej w energię cieplną, np. w trakcie spalania paliwa wprawiającego w ruch tłoki silnika. Tak rozumiane mechanizmy nie pozwalają wyjaśnić zjawisk normatywnych, ponieważ nie ma się tutaj do czynienia z taką asymetrią, jaka powoduje pojawienie się napięcia pomiędzy układem a jego otoczeniem. Nie można zatem stwierdzić, że dane zachowanie jest w jakimś stopniu lepsze bądź poprawniejsze niż inne. Wszystkie zmienne mają taką samą wartość. Jeżeli zatem pojęcie mechanizmu powiążemy z pierwszą zasadą termodynamiki, to krytyka Bickharda wydaje się uzasadniona. Tak jednak nie jest. Jak już wiadomo, pojęcie mechanizmu jest stosowane w wielu różnych naukach, w szczególności w naukach biologicznych i neuronaukach, ale także w kognitywistyce. Gra ono w nich nietrywialną rolę eksplanacyjną (por. Machamer, Darden, Craver, 2011). Nauki te wyjaśniają funkcjonowanie organizmów żywych, czyli takich, do jakich, zdaniem Bickharda i Fristona, stosuje się drugą zasadę termodynamiki, zgodnie z którą układy żywe unikają równowagi termodynamicznej z otoczeniem[47]. Ujmują one zatem relacje asymetryczne, które zdaniem Bickharda są istotne dla wyjaśnienia pojawienia się normatywności. Można zatem stwierdzić, że jego krytyka jest zasadna tylko w odniesieniu do tych dziedzin nauki, które nie odnoszą się do drugiej zasady termodynamiki. Problematyczne jest bowiem zastosowanie pojęcia mechanizmów. Wydaje się, że z perspektywy

[47] Nie wnikam tutaj w różnice pomiędzy teoretyczno-informacyjnym ujęciem Fristona a energetycznym ujęciem Bickharda. Nie jest to istotne dla prezentowanych tutaj analiz.

przeprowadzonych analiz pojęcie to ma odmienne zastosowanie w biologii czy neuronaukach, a inne w klasycznej mechanice[48].

Proponowane tutaj ujęcie sugeruje inne wykorzystanie pojęcia mechanizmu niż proponuje Bickhard. Po pierwsze nie jest dla mnie istotne odwołanie się do zasad termodynamiki, ponieważ, za czym argumentowałem w rozdziale *Zasada energii swobodnej w przetwarzaniu predykcyjnym*, budowanie zasad wyjaśniających zjawiska biologiczne na podstawie pojęć fizyki teoretycznej budzi wiele wątpliwości i prowadzi do słabych eksplanacyjnie wyjaśnień. Po drugie pojęciem wyjściowym dla wyjaśnienia normatywności na tym poziomie badań nie jest, jak u Bickharda, pojęcie funkcji, ale pojęcie mechanizmu. Mechanizmy z jednej strony są normatywne, jeżeli realizują funkcje normatywne. Ale normatywność funkcji nie jest pierwotna względem normatywności mechanizmów. Innymi słowy, jeżeli funkcja reprezentacyjna jest normatywna, to tylko z tego powodu, że istnieje odpowiedni mechanizm, który umożliwia jej pojawienie się. Oznacza to, że predykat „normatywny" w pewnym sensie przypisuje się w tym przypadku wpierw funkcjom, a wtórnie mechanizmom, ale owo przypisanie jest możliwe tylko o tyle, o ile istnieje mechanizm realizujący daną funkcję. Z drugiej jednak strony mechanizm jest normatywny wtedy, kiedy może realizować daną funkcję, a w danym momencie jej nie realizuje. Mechanizm skurczu i rozkurczu komór serca jest normatywny dla funkcji doprowadzania tlenu do komórek organizmu, nawet jeżeli np. z powodu zawału serca w danym momencie tej funkcji nie realizuje. Należy zatem stwierdzić, wbrew Bickhardowi, że o normatywności funkcji nie decyduje ich odniesienie do napięcia pomiędzy asymetrią zachodzącą na linii organizm – środowisko, ale wewnętrzna organizacja systemu. Praca układu krążenia nie wymaga odwołania się do interakcji pomiędzy organizmem a jego otoczeniem. Inaczej jest oczywiście w przypadku funkcji reprezentowania czy funkcji predykcyjnej. Tutaj już trzeba zgodzić się z Bickhardem, że w tych przypadkach nie wystarczy odwołanie się do struktury systemu. Istotna jest bowiem relacja ze środowiskiem, dzięki której reprezentacje, *resp.* predykcje, zyskują warunki prawdziwościowe. Jednak w przeciwieństwie do Bickharda twierdzę, że owe funkcje wyłaniają się z mechanizmów normatywnych, a dla ich wyjaśnienia konieczne jest wzięcie pod uwagę określonych warunków otoczenia. Nie potrzeba jednak odwoływać się do podstaw fizyki teoretycznej, aby wyjaśnić emergencje normatywności ze zjawisk nienormatywnych. Wielopoziomowa i hierarchiczna architektura mechanizmu jest wystarczająca. Zgodnie z poglądem neomechanicystów, chcąc wyjaśnić dany mechanizm, wskazuje się na sposób jego realizacji w mechanizmie leżącym poniżej. Przykładowo, jeżeli chce się wyjaśnić mechanizm pompowania tlenu do komórek przez serce, powinno

[48] Otwartą kwestią jest problem tego, czy neomechanicyzm pozwala na wykrycie takich asymetrii w różnych zastosowaniach pojęcia mechanizmu. W tym kontekście interesujące są uwagi Bechtela (2019), który stwierdza, wbrew wcześniejszym ujęciom, że w wyjaśnieniach mechanistycznych należy również uwzględnić energię (por. także Moreno, 2016). W tym sensie mechanicyzm zbliża się do stanowiska Bickharda.

się odwołać do mechanizmu skurczu i rozkurczu komór serca. Nie trzeba jednak odwoływać się np. do mechanizmu spalania glukozy przez komórki. Mechanicyzm, pomimo odwoływania się do poziomu organizacji mechanizmów, nie jest ujęciem redukcjonistycznym. To, do jakiego mechanizmu odwołujemy się w wyjaśnianiu, jest podyktowane praktyką badawczą i strategią eksplanacyjną.

Powyższe analizy prowadzą do stwierdzenia, że wyjaśnienia wymagają przede wszystkim mechanizmy, a nie funkcje. Oznacza to, że problem emergencji funkcji normatywnych (Bickhard, 2014) zostaje sprowadzony do problemu możliwości pojawienia się mechanizmów normatywnych. Zatem to mechanizmy są emergentne, a nie funkcje. W odniesieniu do PP można powiedzieć, że istnieją takie poziomy modelu generatywnego, jakie realizują odpowiednie funkcje normatywne, oraz takie, jakie ich nie realizują. Odpowiedź na pytanie o to, jakie to są dokładnie poziomy, uzależniona jest od konkretnych badań i jest raczej kwestią rozstrzygnięć empirycznych, co oznacza, że należy je empirycznie odkryć, a nie wydedukować (por. Miłkowski, 2015a, s. 710). Moim zadaniem jest określenie apriorycznych warunków, które mają spełniać mechanizmy, jeżeli mamy o nich twierdzić, że są normatywne. Można już teraz podejrzewać, że normatywne są te poziomy modelu generatywnego, które wiążą się z rozpoznawaniem wzorców, percepcją postaci, kontrolą działania oraz podejmowaniem decyzji[49].

Konkludując, w przeciwieństwie do Bickharda twierdzę, że mechanicyzm jest odpowiednią ramą eksplanacyjną dla badania zjawisk normatywnych ujmowanych z perspektywy biologicznej. Istotne bowiem jest pokazanie, jakie dokładnie role przyczynowe grają odpowiednie funkcje normatywne. Prezentowane badania oczywiście ujmują funkcje normatywne abstrakcyjnie w taki sposób, aby mogły być one realizowane przez różne mechanizmy, niemniej określają one warunki konieczne, choć niewystarczające, do bycia funkcją czy mechanizmem normatywnym. Określam tutaj zatem możliwość pojawienia się takich, a nie innych mechanizmów i funkcji normatywnych, nie wskazując dokładnie biologicznych i fizycznych podstaw ich realizacji. Ostatnim krokiem w tym rozdziale będzie określenie warunków pozwalających na przypisanie normatywności, tj. funkcji prawdziwościowych określonym predykcjom. Konieczne będzie zatem pokazanie, wbrew temu, co twierdzi wielu badaczy (por. np. Gładziejewski, 2016, 2017b; Hohwy, 2015a), że wyjaśnienie określonego normatywnego mechanizmu predykcyjnego wiąże się ze wskazaniem takiego eksplanansu, jaki nie tylko powołuje się na określone stany komponentów ośrodkowego układu nerwowego czy jakiegoś stanu wewnętrznego organizmu, ale także na faktyczne stany środowiska zewnętrznego, w którym dany organizm żyje i funkcjonuje.

[49] W rozdziale *Mechanizmy normatywne a działania w przetwarzaniu predykcyjnym* pokazuję, w jaki sposób „działa" proponowany tutaj model normatywności.

4.5.2.3. Wyjaśnienie mechanizmów normatywnych a ograniczenia

W rozdziale *Przetwarzanie predykcyjne jako bayesowski model eksplanacyjny* (por. podrozdział *Poziomy analizy a przetwarzanie predykcyjne*) omawiałem trzy poziomy analizy wskazane przez Marra (1982): obliczeniowy, algorytmiczny i implementacji. Pierwszy z nich odpowiada na pytanie o to, *co* dany system robi; drugi wskazuje na to, w *jaki* sposób system robi to, co robi; ostatni zaś dotyczy środowiska fizycznego, chemicznego, biologicznego lub społecznego, w którym realizowany jest dany algorytm. Dla przeprowadzonych tutaj analiz kluczowy jest poziom algorytmiczny, ponieważ pozwala on określić, jakie komponenty, działania oraz organizacja systemu pozwalają na realizację mechanizmów normatywnych. Aby móc opisać zjawisko na tym poziomie, należy określić dane na wejściu i wyjściu oraz wskazać na sposób realizacji obliczeń. Zgodnie z teorią wyjaśnień mechanistycznych trzeba zatem pokazać sposób, w jaki (1) jego zmienne odpowiadają komponentom, działaniom, właściwościom i organizacji wyjaśnianego mechanizmu, który jest odpowiedzialny za realizację danego zachowania czy funkcji; oraz (2) wykazać te zależności pomiędzy zmiennymi w modelu, które odpowiadają relacjom przyczynowym zachodzącym pomiędzy komponentami wyjaśnianego mechanizmu (Kaplan, Craver, 2011, s. 611). W odniesieniu do bronionej tutaj tezy, są to dane komponenty i działania systemu oraz określone stany środowiska. W jaki sposób owe stany środowiska włączyć do algorytmu, który ma wyjaśnić sposób realizacji mechanizmu? W niniejszych analizach zaproponuję szkicową odpowiedź na to pytanie, ograniczając się tym samym tylko do pewnego ogólnego modelu mechanizmu.

Marr podkreślał, że jeżeli dane zachowanie (określona funkcja mechanizmu[50]) jest zasadne w odniesieniu do wykonywanego zadania (np. minimalizacji takiego, a nie innego błędu predykcyjnego), to można je uznać za trafne lub adekwatne (*appropriateness*) (Marr, 1982, s. 23–24). Oron Shagrir (2010), komentując tę tezę Marra, stwierdza, że określone zachowanie, które chcemy wyjaśnić, można uznać za trafne lub adekwatne wtedy, gdy jego matematyczny opis da się zmapować na (faktyczną lub kontrfaktyczną) właściwość środowiska, w którym działa dany system. Co to znaczy? Shagrir analizuje przykład wykrywania krawędzi. W tym przypadku proces mapowania, który rozpoczyna się od obrazu na siatkówce, jest „wewnętrznym" procesem mechanicznym, który jest realizowany przez mózg. Jednak opisane na poziomie obliczeniowym zadanie, jakim jest przetwarzanie informacji, jest definiowane, przynajmniej częściowo, przez „zewnętrzne" cechy w polu widzenia, takie jak odbicie światła, oświetlenie i granice obiektów. Wyjaśnienie zatem nie polega tylko na opisie procesu mapowania, ale także na wykazaniu, dlaczego

[50] Wcześniej określiłem funkcje jako działania, za pośrednictwem których przedmioty przyczyniają się do pracy mechanizmu. W tym sensie można je traktować jako rodzaje przyczyn, wyjaśniające np. określone zachowania systemu żywego.

ten proces ma się kończyć na przedstawieniu fizycznych krawędzi, a nie np. na przedstawieniu koloru (Shagrir, 2010, s. 487)? Aby odpowiedzieć na to pytanie, Marr wprowadził pojęcie fizycznych ograniczeń (*physical constraints*). Zadanie wskazywania i wydzielania ograniczeń fizycznych bądź naturalnych umożliwia zdefiniowanie określonych procesów. Owe ograniczenia pozwalają sądzić, że procesy, które je uwzględniają, są prawdziwe (Marr, 1982, s. 23). W odniesieniu do procesów wykrywania krawędzi naturalnym ograniczeniem jest lokalizacja przestrzenna (Marr, 1982, s. 68–70). Oznacza to, że przedmioty w świecie, który powoduje zmiany intensywności światła, są zlokalizowane przestrzennie. Należy zatem stwierdzić, że wyjaśnienie procesu wykrywania krawędzi powinno również objąć, jako potencjalny komponent tego mechanizmu, ograniczenia fizyczne. Przez owe ograniczenia można rozumieć określone fakty dotyczące świata realnego (Shagrir, 2010, s. 489). Shagrir podkreśla również, że wyjaśnienie sformułowane przez Marra odwołuje się do podobieństwa pomiędzy wewnętrznymi relacjami mapowania i zewnętrznymi relacjami zachodzącymi pomiędzy reprezentowanymi cechami. Podobieństwo nie dotyczy jednak poziomu właściwości fizycznych, bowiem ostatecznie fizjologiczne własności mózgu są zupełnie inne niż własności fizyczne i optyczne środowiska, czyli takie, jakie składają się na nasze pole widzenia. Podobieństwo występuje na bardziej abstrakcyjnym poziomie właściwości matematycznych. Uwaga ta jest kluczowa dla prowadzonych przeze mnie analiz, ponieważ podkreśla ona konieczność uwzględnienia w wyjaśnieniach określonych stanów środowiska. Marr pokazuje, że proces wykrywania krawędzi należy wyjaśnić nie tylko poprzez wskazanie wewnętrznej relacji mapowania pomiędzy siatkówką oka a określonymi populacjami neuronów w mózgu (rejon V1), ale równie ważne jest opisanie relacji pomiędzy natężeniami światła w polu widzenia a nagłymi zmianami natężenia światła w otoczeniu. Zmiany te bowiem zdarzają się, przynajmniej w naszym świecie, wzdłuż granic obiektów. Oznacza to, że istnieje określona zależność matematyczna pomiędzy reprezentowanymi cechami, np. pomiędzy odbijaniem światła w polu widzenia i odbijaniem światła wzdłuż granic spostrzeganego obiektu. Marr zatem nie tylko pokazuje, że wewnętrzna funkcja matematyczna koreluje ze światem przygodnym, w którym żyjemy, ale także, że jej podstawą jest podobieństwo struktur matematycznych. Zdaniem Shagrira owo podobieństwo jest kluczowe dla adekwatności wyjaśnień budowanych na poziomie obliczeniowym (Shagrir, 2010, s. 489). Jeżeli nie uchwyci się tych relacji w wyjaśnieniu, to co prawda można stworzyć określony model eksplanacyjny danego zjawiska. Będzie to jednak model zjawiska abstrakcyjnie ujętego, czyli w pewnym sensie wyjaśnionego poza kontekstem środowiskowym i funkcjonalnym.

Marr, wyjaśniając pojęcie ograniczenia, podał przykład robienia zakupów (Marr, 1982, s. 22). Mając określoną ilość pieniędzy, mogę zakupić wiele różnych artykułów. Aby wyjaśnić, na czym polega proces robienia zakupów, należy wykazać, dlaczego dodawanie liczb jest odpowiednio zdefiniowanym algorytmem dla

przetwarzania informacji w tym kontekście, czyli w sklepie spożywczym. Można np. zapytać: „dlaczego kasa fiskalna dodaje ceny produktów, a nie przykładowo je mnoży, aby otrzymać ostateczny rachunek?" (Marr, 1982, s. 22). Nie ma przecież powodu *a priori*, aby ostatecznym rachunkiem była suma pieniędzy wydana na zakupione artykuły, a nie ich iloczyn. Powodem jest to, twierdzi Marr, że ograniczenia robienia zakupów w sklepie definiują zasady dodawania. Owe ograniczenia Marr nazywa również regułami (*rules*) kupowania artykułów w sklepie. W tym przypadku można wyróżnić cztery takie reguły: (1) zerowa – „jeśli nic nie kupisz, nie powinno cię to nic kosztować"; (2) przemienności – „kolejność, w jakiej towary są przedstawione kasjerowi nie powinna wpływać na sumę rachunku"; (3) łączności – „ układanie towarów w dwa stosy i płacenie za każdy stos osobno nie powinno wpływać na całkowitą kwotę, którą płacisz"; oraz (4) odwrotności – „jeżeli kupisz przedmiot, a następnie zwrócisz w go celu uzyskania zwrotu, całkowite wydatki powinny wynosić zero" (Marr, 1982, s. 22–23). Owe reguły są więc normami definiującymi to, co nazywamy „robieniem zakupów". Są one arbitralne w tym sensie, że można z łatwością wymyślić inny sposób kupowania przedmiotów, nie są jednak względne, ponieważ albo się ich przestrzega albo nie. Jeżeli się ich nie przestrzega, to nie robi się zakupów, tylko przykładowo kradnie artykuły.

Ogólnie mówiąc, tak rozumiane ograniczenia są w pewnym sensie normami, określającymi granice i zasady realizacji danych procesów. Dzięki nim badany proces bądź mechanizm jest realizowany w określonym środowisku przez odpowiedni system je realizujący. W tym sensie są one ograniczeniami normatywnymi i jako takie powinny być brane pod uwagę w wyjaśnieniu funkcjonowania określonych mechanizmów. Są one bowiem przyczynowo skuteczne: zapewniają niezbędne i wystarczające warunki dla funkcjonowania określonych procesów (Marr, 1982, s. 111–116)[51]. Przykładowo w procesie widzenia gwarantują dopasowanie odpowiednich elementów do większości scen naturalnych (por. Marr, Poggio, 1976, 1979). Pojęcie to jest kluczowe dla analizy normatywności mechanizmów predykcyjnych i mechanizmów normatywnych w ogóle. Jego pełną eksplikację przeprowadzam w rozdziale *Mechanizmy normatywne a działania w przetwarzaniu predykcyjnym*. Na ten moment należy powiedzieć, że wyjaśnienie mechanizmów normatywnych, w tym mechanizmów predykcyjnych, powinno obejmować opis komponentów oraz relacji, w jakie one wchodzą, ich działań oraz takich ograniczeń fizycznych, jakie wspólnie są odpowiedzialne za pojawienie się danego zjawiska. Istotne jest to, że nie wszystkie zjawiska dają się wyjaśnić w kategoriach mechanizmów neuronalnych (por. Weiskopf, 2016)[52]. Czasem należy powołać się na odpowiednie komponenty i operacje, które są współkonstytuowane także przez ograniczenia społeczne i kulturowe (Miłkowski

[51] Uczniowie Marra stwierdzają, że teoria obliczeniowa obejmuje „analizę tego, w jaki sposób właściwości świata fizycznego ograniczają rozwiązywanie problemów w widzeniu" (Hildreth, Ullman, 1989, s. 582).

[52] Na temat rodzajów i zastosowania pojęcia ograniczenia w mechanicyzmie por. Winning, 2018.

i in., 2018, s. 9; Norman, 2018). Należy jednak nieustannie pamiętać o podstawowej normie wyjaśnień mechanistycznych, zgodnie z którą adekwatne i nietrywialne wyjaśnienie to takie, jakie dobrze odwzorowuje rzeczywistą strukturę przyczynową mechanizmu odpowiedzialnego za wyjaśniane zjawisko (por. Kaplan, 2011[53]).

Podsumowanie

W niniejszym rozdziale zostały zdefiniowane pojęcia funkcji i mechanizmu normatywnego. Argumentowałem za mechanistycznym odczytaniem tych pojęć, opierając się przy tym na rolach przyczynowych, które te pojęcia odgrywają w potencjalnych wyjaśnieniach zjawisk poznawczych. Pokazałem również, że satysfakcjonujące i nietrywialne wyjaśnienie normatywnych mechanizmów predykcyjnych wymaga odniesienia się nie tylko do komponentów, operacji oraz działań wykonywanych przez dany mechanizm, ale także do określonych stanów środowiska zewnętrznego, które wiążą się z istnieniem tzw. ograniczeń dla mechanizmu.

Celem prezentowanych analiz nie było pokazanie, w jaki sposób realizowane są fizycznie i materialnie normatywne mechanizmy predykcyjne, a uzasadnienie, że są one istotne dla analiz przeprowadzanych w ramach PP, oraz, że tak rozumiane mechanizmy mają określoną moc eksplanacyjną w ramach mechanistycznej koncepcji wyjaśnień naukowych. Zostało wykazane, że (1) PP ma zastosowanie na wszystkich trzech poziomach analizy w rozumieniu Marra; (2) zasadne jest przyjęcie stanowiska preskryptywizmu ontycznego w odniesieniu do PP ze względu na ontyczny charakter wyjaśnień formułowanych na gruncie tego podejścia; (3) bayesowski model PP jest nietrywialną perspektywą badawczą ze względu na normatywny charakter mechanizmów bayesowskich. Innymi słowy, wykazałem, że istnieją zjawiska, które wyjaśnia się przy pomocy bayesowskiego PP, ponieważ są one bayesowskie z natury, a jeżeli są one bayesowskie z natury, to znaczy, że powinno się je wyjaśniać przy użyciu modelowania probabilistycznego. Takimi zjawiskami są mechanizmy normatywne, które grają określone role przyczynowe w wyjaśnieniach działań. Mówiąc o takich rolach w odniesieniu do normatywnych własności mechanizmów oraz funkcji, stwierdziłem, że są one przyczynami określonych działań danego organizmu bądź podmiotu w jego środowisku.

[53] Część autorów (por. Gładziejewski, 2019; Harkness, Keshava, 2017; Hohwy, 2015a) twierdzi, że PP dostarcza szkicu mechanizmu. Szkic mechanizmu w rozumieniu Gualtiero Piccininiego i Davida Kaplana (2011) należy zdefiniować w terminach ról funkcjonalnych granych przez odpowiednie komponenty, abstrahując jednocześnie od ich biologicznej bądź fizycznej realizacji. W tym kontekście zasadne jest pytanie dotyczące tego, w jaki sposób rozumieć przyczynową strukturę odpowiedzialną za mechanizmy predykcyjne? Może być to prosta wielopoziomowa hierarchia rozpoczynająca się od prostych poziomów neuronalnych dotyczących rozpoznawania wzorców, wykrywania krawędzi, percepcji koloru itd., a kończąca na coraz bardziej abstrakcyjnych i ogólnych poziomach powiązanych z przekonaniami i dotyczących ogólnych własności świata; albo bardziej wysublimowana struktura realizowana przez kilka różnych, w jakimś stopniu niezależnych, mechanizmów odpowiedzialnych za różne zjawiska. Pytanie to pozostawiam otwarte.

Rozdział v
Mechanizmy normatywne a działania w przetwarzaniu predykcyjnym

Wstęp

Rozdział ten jest poświęcony problemowi ciemnego pokoju i powiązanemu z nim kompromisowi eksploracji i eksploatacji. Ciemny pokój oznacza taki stan, w którym mógłby znaleźć się podmiot, gdyby zminimalizował sumę wszystkich potencjalnych błędów predykcyjnych. Wydaje się, że zgodnie z bazowym założeniem PP o konieczności ciągłej i długoterminowej minimalizacji błędów predykcyjnych, taki stan powinien być dla podmiotu pożądany. Czy tak właściwie jest? Wielu autorów sądzi, że nie. Będę twierdził, że sprawdzianem wartości PP oraz w pewnym stopniu PP jest możliwość nietrywialnego rozwiązania tego problemu, który można sprowadzić do wyboru pomiędzy aktywną i zwiększającą niepewność eksploracją a bezpieczną i łatwo przewidywalną eksploatacją. Analizując kolejne propozycje rozwiązania problemu ciemnego pokoju, będę argumentował za tym, że w pewnym stopniu ani (powiązana z FEP) rama aktywnego wnioskowania, ani samo PP nie są w stanie zaoferować satysfakcjonującego wyjaśnienia tego kompromisu. Wykażę, dlaczego tak jest.

W dalszej części rozdziału odniosę się do bronionej tutaj koncepcji mechanizmów normatywnych oraz dokonam szczegółowych analiz, kluczowego dla niniejszych rozważań, pojęcia motywacji. W tym kontekście będę twierdził, że pełne wyjaśnienie mechanizmów normatywnych, a w dalszej kolejności rozwiązanie dylematu eksploracji i eksploatacji zakłada odwołanie się do istnienia ograniczeń obecnych w środowisku. Ograniczenia są tym elementem, który czyni dany mechanizm nie tylko przyczynowym, ale także normatywnym. Kluczowe zatem dla wyjaśnienia mechanizmów predykcyjnych są określone stany środowiska (nie tylko fizycznego, ale także społeczno-kulturowego itd.), które można analizować w kategoriach ograniczeń dla mechanizmów. Owe ograniczenia nie pełnią tylko funkcji kontekstu, w którym mechanizm jest realizowany, ale przede wszystkim są jego konstytutywnym komponentem. Będę dowodził, że pełne wyjaśnienie roli niektórych ograniczeń w normatywnych mechanizmach predykcyjnych zakłada określoną integrację poszczególnych modeli takich, a nie innych zjawisk poznawczych, bowiem dopiero (mechanistyczna) integracja PP (bądź ramy aktywnego

wnioskowania) z innym modelem, *resp.* modelami, pozwoli na dobre wyjaśnienie natury normatywnych mechanizmów predykcyjnych. Będę również bronił tezy, zgodnie z którą niemożliwość rozwiązania problemu ciemnego pokoju przez PP jest bezpośrednim następstwem monizmu eksplanacyjnego przyjętego przez zwolenników tego podejścia.

W ostatniej części tego rozdziału pokażę, że mechanizmy normatywne stanowią fragment większej całości, ponieważ o ich normatywnym charakterze decydują nie tylko jego funkcje odniesione do organizacji i struktury organizmu, do którego stabilności dany mechanizm się przyczynia, ale także odniesienie do określonych stanów środowiska, rozumianych jako komponenty mechanizmu konstytutywne dla wyjaśnienia danego zjawiska. Biorąc pod uwagę aspekt środowiskowy mechanizmów normatywnych, okazuje się, za czym będę argumentował, że stanowią one element pewnej szerszej struktury normatywnej, którą nazwę „normatywnością rozproszoną". Hipoteza normatywności rozproszonej zakłada, że o normatywnym charakterze określonych mechanizmów, funkcji czy przedmiotów decydują relacje, w które te mechanizmy, funkcje czy przedmioty wchodzą. W odniesieniu do tak sformułowanej hipotezy przedyskutuję poglądy zwolenników enaktywistycznego ujęcia PP i wykażę słabość eksplanacyjną tego stanowiska.

5.1. Uwagi wprowadzające

W poprzednim rozdziale zaproponowałem ogólny model normatywności dla mechanizmów i funkcji predykcyjnych w odniesieniu do ramy PP. Wskazałem elementy konieczne, pozwalające mówić o tym, że dany mechanizm bądź funkcja są normatywne. Stwierdziłem przy tym, że satysfakcjonujące i nietrywialne wyjaśnienie normatywnych mechanizmów predykcyjnych wymaga odniesienia się nie tylko do komponentów, operacji oraz działań (funkcji) wykonywanych przez dany mechanizm, ale także do określonych stanów środowiska zewnętrznego, które powiązałem z istnieniem tzw. ograniczeń dla mechanizmów. W niniejszym rozdziale pokażę, w jaki sposób funkcjonuje zaproponowany przeze mnie model normatywności. Mówiąc precyzyjniej: będę chciał uzasadnić tezę, że zachowania i działania usytuowanego w środowisku podmiotu dają się wyjaśnić poprzez odwołanie się do istnienia modelu generatywnego realizującego odpowiednie normatywne mechanizmy predykcyjne. Moja argumentacja będzie wiązała się z tzw. kompromisem pomiędzy badaniem (eksploracja) a wykorzystywaniem (eksploatacja) (*explore-exploit trade-off*). Jest to pojęcie zaczerpnięte z badań nad uczeniem maszynowym, ale ma ono znacznie szersze zastosowanie. Najogólniej mówiąc, kompromis ten dotyczy sytuacji, w których wybiera się pomiędzy tym, co jest znane i czego zastosowanie da się przewidzieć (może spełnić ono nasze

oczekiwania) (eksploatacja), a tym, czego nie jest się pewnym, ale istnieje mocne przypuszczenie, że może ono zaoferować pewną *nowość* – w postaci informacji, doświadczenia czy umiejętności (eksploracja). Eksploracja dostarcza nową wiedzę, tworzy nowe możliwości działań, ale wiąże się z większym ryzykiem, powoduje pojawienie się niepewności[1]. Eksploatacja zaś zmniejsza potencjalne ryzyko, co oznacza, że określone wydarzenia stają się bardziej przewidywalne. Pozwala ona zatem na minimalizację niepewności. Rozważenie kilku prostych przykładów pozwoli lepiej zrozumieć ten kompromis.

Można sobie wyobrazić wieczorne wyjście do restauracji. Czy chcemy wybrać lokal znany i sprawdzony, czy pójść do innego, który może nas pozytywnie zaskoczyć? Czy chcemy dalej pracować w dobrze płatnej, ale już nudnej dla nas pracy, czy zaryzykować i wybrać nową propozycję zatrudnienia? Czy zgubiwszy się w lesie, lepiej najeść się rosnącymi na małych krzaczkach jagodami, co zajmie wiele czasu, czy może lepiej wspiąć się na dziką jabłoń i zerwać wysoko rosnące jabłka, co może skutkować nie tylko szybkim zabiciem głodu, ale także możliwym upadkiem i innymi nieprzyjemnymi konsekwencjami? To, którą z możliwości się wybierze, zależy od tego, jak kosztowne są informacje o określonych konsekwencjach, jak długo będzie można z nich skorzystać oraz jak duże są z nich korzyści.

Należy powiedzieć, że eksploracja jest procesem czaso- i energochłonnym. Wiąże się to z nieobecnością określonych informacji dotyczących danej sytuacji czy przedmiotu. Nigdy do końca nie wiadomo, jak dużo pracy trzeba włożyć w dany proces i które informacje należy wykorzystać. Istnieje jeszcze inny problem. Zbyt dużo eksploracji może prowadzić do nieefektywnych i nieoptymalnych działań, podczas gdy zbyt duże wykorzystywanie danych zasobów czy określonych możliwości może uczynić działania nieefektywnymi i nieprowadzącymi do przewidzianych efektów (por. Jasrasaria, Pyzer-Knapp, 2018). Jest to problem, z którym PP powinno się zmierzyć.

Ze względu na cele postawione w tych analizach zakładam, że kompromis eksploracja – eksploatacja dotyczy (1) wyboru pomiędzy zdobywaniem nowej wiedzy a wykorzystaniem wiedzy już istniejącej; oraz (2) wyboru pomiędzy nowymi nieoczywistymi możliwościami działań a sprawdzonymi i znanymi strategiami działania. Na ten moment należy stwierdzić, że z eksploatacją wiąże się minimalizowanie poziomu niepewności. Niemniej może być tak, że korzyść, jaką można zyskać, wybierając eksploatację, może być relatywnie niższa niż w przypadku

[1] Niepewność wiąże się z takimi sytuacjami, w jakich określone decyzje mogą prowadzić do różnych skutków w zależności od tego, który z możliwych stanów rzeczy zajdzie, przy czym nie są znane prawdopodobieństwa wystąpienia poszczególnych z nich. Ryzyko zaś dotyczy takich sytuacji, w których podjęcie odpowiedniego działania pociąga za sobą możliwości wystąpienia różnych negatywnych i pozytywnych konsekwencji przy znanym prawdopodobieństwie wystąpienia każdej możliwości. W literaturze rozróżnia się pojęcia ryzyka i niepewności. Jednak większość decyzji podejmowana jest na środkowym polu pomiędzy znanymi a nieznanymi prawdopodobieństwami. Z tego powodu trudno oddzielić decyzje w warunkach niepewności od tych podejmowanych w warunkach ryzyka.

wyboru eksploracji. W tej ostatniej zaś niepewność się zwiększa, niemniej i potencjalne korzyści mogą być większe. Rozróżnienia te zyskają niebawem na jasności.

Twierdzę, że właściwą ramą dla badania kompromisu eksploracja – eksploatacja w PP stanowi tzw. problem ciemnego pokoju (*dark room problem*). Wiąże się on z ogólnym zarzutem stawianym PP, a dotyczącym głównego założenia tego podejścia, zgodnie z którym optymalna i konieczna dla podmiotów poznających jest ciągła minimalizacja niepewności. Ogólnie mówiąc, zarzut ten polega na tym, że według bazowych założeń PP działający podmiot powinien chcieć znaleźć się w ciemnym, nieoferującym żadnych bodźców pokoju, ponieważ taki stan pozwala mu w sposób ciągły i kontrolowany minimalizować błędy predykcyjne, czyli niepewność. W tych warunkach każda nowa sytuacja jest przewidywalna, zatem poziom zaskoczenia jest bliski zeru. Twierdzę, że nietrywialna odpowiedź na problem ciemnego pokoju stanowi o zasadności i nietrywialności określonego modelu PP. Uważam również, że proponowana w tej pracy koncepcja normatywności mechanizmów predykcyjnych pozwoli wyjaśnić, dlaczego PP jest odporne na różne krytyki widzące w problemie ciemnego pokoju paradoks, z którym ta rama badawcza nie umie sobie poradzić.

Zanim jednak przejdę do szczegółowego omówienia problemu ciemnego pokoju, konieczne są dodatkowe ustalenia. Wielu badaczy oferuje różne wyjaśnienia z zakresu PP dotyczące możliwości działania, kontroli motorycznej, eksploracji środowiska czy wartości epistemicznych oraz nieepistemicznych oferowanych przez otoczenie. Różnią ich jednak nie tylko szczegółowe rozwiązania i sugestie badawcze, ale także oczekiwania eksplanacyjne. Mam na myśli pewne oczekiwania co do możliwości eksplanacyjnych, jakie przypisuje się PP, a także wiązanej z nim przez niektórych autorów FEP.

Andrew Sims (2017) zaproponował przegląd różnych strategii eksplanacyjnych obecnych wśród zwolenników PP i FEP. Jako pierwsze wyróżnił stanowisko minimalnego PP. Jego zakres eksplanacyjny obejmuje tylko niektóre procesy percepcyjne. Dobrym przykładem tego podejścia jest praca Rao i Ballarda (1999; por. podrozdział *Możliwe algorytmy przetwarzania predykcyjnego i filozoficzne inspiracje*). Następnie mamy stanowisko umiarkowanego (*mixed*) PP. Zgodnie z tym ujęciem PP może wyjaśnić niektóre procesy percepcyjne i motoryczne. W ujęciu maksymalnego PP wyjaśnieniu podlegają wszystkie procesy neuropoznawcze. Na koniec Sims wyróżnia stanowiska, które zakładają prawdziwość FEP. W tych ujęciach FEP (według niektórych autorów) skorelowane z PP pozwala wyjaśnić wszelkie procesy biologiczne, w różnych skalach czasowych (Sims, 2017, s. 5). W zależności od tego, którą z wyróżnionych strategii się przyjmie, to osiągnie się takie, a nie inne rezultaty poznawcze. Jest to istotne, bowiem szczególnie ci badacze, którzy uznają normatywny i eksplanacyjny wymiar FEP, są skłonni tłumaczyć większość, jeżeli nie wszystkie zachowania i działania podmiotów, poprzez odniesienie do wymogu minimalizacji energii swobodnej. W rozdziałach *Zasada*

energii swobodnej w przetwarzaniu predykcyjnym i *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego* argumentowałem przeciwko tej tezie. Twierdzę, że FEP nie jest zasadą eksplanacyjną, a pewną ogólną idealizacją, którą można potraktować jako heurystykę badawczą. Z tego też powodu odrzucam pogląd, głoszony przez zwolenników FEP, że przy pomocy PP można wyjaśnić wszelkie procesy biologiczne i poznawcze, włącznie z tzw. wyższymi procesami poznawczymi (*higher cognition*), tj. rozumowaniem abstrakcyjnym, myśleniem pojęciowym, sądzeniem itd.[2]

W niniejszej pracy przyjmuję pewną wersję poglądu umiarkowanego, ponieważ uważam, że rama badawcza PP pozwala wyjaśnić tę część z naszych procesów percepcyjnych i poznawczych, która jest ściśle skorelowana z działaniami w środowisku. Jest to jednak szczególna wersja tego poglądu, bowiem zakładam konieczność przyjęcia pluralizmu eksplanacyjnego. Mianowicie: dopiero PP zintegrowane z innymi modelami może efektywnie wyjaśniać określone zjawiska poznawcze.

5.2. Problem ciemnego pokoju

Pierwsze sformułowanie problemu ciemnego pokoju zostało podane przez Davida Mumforda w pochodzącym z 1992 roku artykule pt. *On the computational architecture of the neocortex*. Mumford (1992, s. 247) powiada:

> W pewnym sensie jest to stan (znajdywania się w ciemnym pokoju – MP), który kora mózgowa stara się osiągnąć: doskonałe przewidywanie świata, coś jak stan nirwany… kiedy nic Cię nie zaskakuje, a nowe bodźce powodują najdrobniejsze falowanie w twojej świadomości.

Sam Mumford nie potraktował tej uwagi jako jakiejś obiekcji, raczej jako pewne spostrzeżenie. W ujęciu wielu krytyków problem ciemnego pokoju jest, nie bez powodu, traktowany jako zarzut (por. Baltieri, Buckley, 2019; Klein, 2018). Również sami zwolennicy PP dostrzegają, że problem ten wymaga rozwiązania.

W interesującym mnie aspekcie problem ciemnego pokoju można wyrazić w następujący sposób: jest to stan, w którym mógłby znaleźć się podmiot, gdyby zminimalizował sumę wszystkich potencjalnych błędów predykcyjnych. Wyobraźmy sobie, że znajdujemy się w cichym, ciemnym pomieszczeniu, w którym nic się nie wydarza. W takich warunkach potrafię bardzo dobrze przewidzieć, co się stanie. Żadne stany świata nie mogą mnie zaskoczyć. Całkowity brak bodźców zapewnia optymalne i maksymalnie wydajne minimalizowanie wszelkiej niepewności. Wydaje się, że zgodnie z bazowym założeniem PP o konieczności ciągłej

[2] Jest wiele nowszych prac zajmujących się tym zastosowaniem FEP. Por. np. Hesp i in., 2019; Kiverstein, Miller, Rietveld, 2019.

i długoterminowej minimalizacji błędów predykcyjnych, taki stan powinien być dla podmiotu pożądany. Ale czy tak rzeczywiście jest? Clark (2013b, s. 191) już w programowym tekście z 2013 roku wyraził tę wątpliwość:

> W jaki sposób neuronalny imperatyw nakazujący minimalizowanie błędów predykcyjnych poprzez zniewolenie (*enslaving*) percepcji, działania i uwagi, uwzględnia oczywisty fakt, że zwierzęta nie szukają po prostu ładnego ciemnego pokoju i nie pozostają w nim? Z pewnością ciągłe przebywanie w zaciemnionym pokoju zapewniłoby łatwą i niemal idealną prognozę naszych rozwijających się stanów neuronalnych.

W innym miejscu stwierdza: „Nieszczęsny organizm napędzany prognozami powinien po prostu szukać stanów, które są łatwe, przewidywalne, takie jak pusty zaciemniony pokój, w którym mógłby spędzić resztę swoich coraz bardziej głodnych, spragnionych i przygnębiających dni" (Clark, 2016, s. 262). Schwartenbeck z Fristonem dodają: „Czy nie powinniśmy, zgodnie z zasadą (FEP – MP), chcieć mieszkać w wysoce przewidywalnym i pozbawionym stymulacji środowisku, w którym moglibyśmy zminimalizować nasze długoterminowe zaskoczenie?" (Schwartenbeck i in., 2013, s. 1).

Problem ciemnego pokoju opisuje wyraźny kontrast pomiędzy bogatym repertuarem zachowań prawdziwych żywych stworzeń a wymogiem sprowadzania wszystkich zachowań, działań i decyzji podmiotu uzbrojonego w model generatywny do wymogu minimalizacji błędów predykcyjnych, *resp.* zaskoczenia i niepewności. Nie jest przecież tak, że żywe, ucieleśnione i osadzone w środowisku stworzenia zawsze poszukują niszy, która spełniałaby wymogi ciemnego pokoju. Raczej jest tak, że środowisko (nie tylko fizyczne, ale także kulturowo-społeczne) i inne podmioty motywują nas do podejmowania różnego rodzaju działań, również tych, które nie minimalizują niepewności, ale przeciwnie zwiększają ją. Czy zatem PP, jako teoria wyjaśniająca, jeżeli nie wszystkie, to przynajmniej część procesów oraz zjawisk percepcyjnych i poznawczych, nie potrafi zdać sprawy z tego, że istnieją takie działania i zachowania, jakie nie zawsze prowadzą do minimalizacji błędów predykcyjnych? A może jednak jest tak, że każde działanie, jakie podejmujemy, ostatecznie polega na tym, że w jakimś stopniu pozwala ono na długoterminową minimalizację zaskoczenia? Na te pytania będę chciał w niniejszym rozdziale odpowiedzieć. Są one kluczowe, ponieważ odpowiedź na nie – jak twierdzę – wymaga odniesienia się nie tylko do wewnętrznej architektury i dynamiki modelu generatywnego i sieci bayesowskiej, ale także do uwarunkowań i ograniczeń obecnych w środowisku zewnętrznym. Bowiem środowisko nie jest tylko źródłem sygnału zmysłowego, ale także, a może przede wszystkim, *miejscem*, w którym pojawiają się i mogą pojawić odpowiednie wartości dla podmiotu. Takie wartości, jakie motywują go do podjęcia odpowiednich działań, ale również takie, jakie powodują, że podmiot pewnych działań nie chce wykonać. Zanim sformułuję własną, opartą na proponowanym w tej pracy modelu

normatywności, odpowiedź, konieczne jest (1) prześledzenie niektórych zarzutów wobec PP, sformułowanych w perspektywie problemu ciemnego pokoju, oraz (2) przedstawienie rozwiązań, jakie zwolennicy PP zaproponowali, aby odeprzeć te krytyki lub rozwiązać dylemat, który bezpośrednio wiąże się ze wspomnianym problemem.

5.2.1. Odpowiedzi zwolenników przetwarzania predykcyjnego

W niniejszych rozważaniach przenalizuję niektóre z odpowiedzi, jakie sformułowali zwolennicy PP i FEP wobec problemu ciemnego pokoju. Następnie rozważę głosy krytyczne wobec tych propozycji i zastanowię się, czy dotychczasowe próby rozwiązania dylematu, który jest implikowany przez problem ciemnego pokoju, są satysfakcjonujące. W ten sposób będę mógł odnieść się do anonsowanego wcześniej kompromisu pomiędzy eksploracją a eksploatacją, który wyznacza analizy dla niniejszego rozdziału.

W jednym z pierwszych tekstów poświęconych problemowi ciemnego pokoju Friston, Thornton i Clark (2012) stwierdzają, że ciemny pokój nie jest zupełnie fikcyjnym pomysłem, ponieważ istnieją troglofile, czyli stworzenia (*dark-room agents*), które są ewolucyjnie dostosowane do życia i nawigacji w zaciemnionych miejscach, takich jak jaskinie, opuszczone bunkry czy wody podziemne. Można zatem z perspektywy omawianego tutaj kompromisu zapytać o to, dlaczego takie zwierzęta istnieją? Odpowiedź jest prosta. Każdy żywy organizm dąży do takiej minimalizacji energii swobodnej, *resp.* niepewności, jaka polega na zmianie sygnału zmysłowego w ten sposób, że realizowane są działania, które odpowiadają ich przewidywaniom i modelowi świata. W tym sensie każdy taki organizm stanowi realne (tj. ewolucyjne) rozwiązanie problemu minimalizacji zaskoczenia w zmieniającym się świecie. Inne są, ukształtowane ewolucyjnie i rozwojowo, architektura oraz sposób optymalizacji działań w środowisku u troglofili, a inne u ludzi. Dlatego też ludzie nie żyją w „ciemnych pokojach", a troglofile tak. Oznacza to zatem, że ciemny pokój zapewnia niski poziom zaskoczenia tylko wtedy, gdy podmiot został tak zoptymalizowany przez ewolucję, aby mógł w nim przebywać i działać (Ramstead, Kirchhoff, Friston, 2019, s. 5). Inne podmioty nie będą chciały takiego pomieszczenia zamieszkiwać i je opuszczą. Podmiot może zatem wybierać optymalne miejsce dla działania i regulowania swojego modelu.

Można zapytać o to, co się jednak dzieje, kiedy podmiot, taki jak człowiek, znajdzie się w ciemnym pokoju? Czy jednak nie jest tak, że będąc w takim pomieszczeniu, zmniejszamy poziom niepewności i zaskoczenia?[3] Zdaniem Clarka nie, ponieważ zwierzęta takie jak my, żyją i funkcjonują w świecie, który się wciąż

[3] Sims (2017) proponuje, aby pytania związane z ciemnym pokojem sprowadzić do następującego: jeżeli działania są jedynie minimalizacją zaskoczenia, to dlaczego nie próbujemy zminimalizować wszystkich możliwych bodźców?

zmienia i wymaga czegoś od nas (por. Clark, 2018). Oczekujemy zatem, że nieustannie coś się będzie działo: „zmiana, ruch, eksploracja i poszukiwanie są same w sobie cenne dla stworzeń żyjących w światach, w których zasoby są nierównomiernie rozmieszczone i stale pojawiają się nowe zagrożenia i szanse" (Clark, 2013b, s. 193). Trudno się z Clarkiem nie zgodzić, jednak jego odpowiedź rodzi pytania, które zadałem we wstępie do tego rozdziału, a które dotyczą mechanizmu regulującego wybór pomiędzy eksploatacją (ciemny pokój) a eksploracją (wyjście z ciemnego pokoju). To, że wiemy, iż wyjdziemy, nie oznacza, że mamy na to dobre wyjaśnienie. Chodzi zatem o wykładnię PP, która takie wyjaśnienie zaoferuje, nie negując jednocześnie postulatu mówiącego o konieczności minimalizacji błędów predykcyjnych.

Inną odpowiedź formułuje Friston. Jego zdaniem pierwszą rzeczą, którą zrobimy po wejściu do ciemnego pokoju będzie włączenie światła. Uważa on, że takie działanie oferuje coś, co określa mianem afordancji epistemicznej (*epistemic affordance*). Chodzi o to, że włączenie światła jest takim działaniem, które minimalizuje niepewność (tj. oczekiwaną energię swobodną). Mianowicie: nie jest tak, że wchodząc do ciemnego pomieszczenia, oczekujemy, że będzie w nim panował mrok. Raczej oczekujemy, że będzie w nim jasno. Zatem minimalizacja niepewności wiąże się z oczekiwaniami, że ciemny pokój może zostać oświetlony, a nie że będzie pozostawał w ciemności (Friston, Fortier, Friedman, 2018, s. 26). „Ludzie nie szukają ciemnych pomieszczeń – tj. środowiska, które jest doskonale płynne. Takie jest zbyt nudne. Szukają raczej słodkich punktów, które charakteryzują się zarówno pewną płynnością (pewność i zażyłość), jak i grudkowatością (niepewność i nieznajomość)" (Friston, Fortier, Friedman, 2018, s. 25). Odpowiedź Fristona, podobnie jak ta sformułowana przez Clarka, raczej odsłania kompromis, który nas interesuje, aniżeli go wyjaśnia.

Giovanni Pezzulo (Pezzulo, Rigoli, Frison, 2015, s. 32) sugeruje, że ujęcie żywych organizmów z perspektywy FEP (ujęcie homeostatyczne) wyjaśnia, dlaczego podmioty nie mają problemu z „ciemnym pokojem". Jego zdaniem regulacja homeostatyczna implikuje ciągłą aktualizację empirycznych priorytetów działania, których dynamika jest podyktowana nieprzerwanym strumieniem komunikatów interoceptywnych przepływających pomiędzy mózgiem a ciałem. Ten strumień nieustannie dostarcza informacji o nowych wartościach oraz celach. Zatem każdy organizm wyposażony w ciało musi stawić czoła rzeczywistości, a nie żyć w zaciemnionym pokoju. Pogląd ten znacznie lepiej tłumaczy powody unikania przez podmioty ciemnych pomieszczeń, niemniej nie wyjaśnia zupełnie, w jaki sposób coś obecnego w środowisku staje się dla nich wartością albo celem. Innymi słowy, jaki mechanizm czyni dla podmiotu środowisko nośnikiem wartości? Dlaczego postrzega je jako mające pewne znaczenie?[4]

[4] Abstrahuję teraz od faktu, że tylko nieliczne z mechanizmów realizowanych przez żywe organizmy można uznać za faktycznie homeostatyczne. Byłbym bardzo ostrożny w traktowaniu mechanizmów, o których mowa, jako homeostatycznych. Por. podrozdział *Normatywność homeostazy*.

Schwartenbeck (Schwartenbeck i in., 2013) podkreśla, że (większość) podmiotów przebywanie w ciemnym pokoju uznałoby za bardzo zaskakujące, zatem starałoby się uniknąć takiego stanu rzeczy. Problem polega na tym, że myli się potoczne znaczenie pojęcia intuicji z pojęciem zaskoczenia z teorii informacji, które należy rozumieć w kategoriach nieprzewidywalnego wkładu sensorycznego i zaskoczenia w odniesieniu do określonego modelu świata. Znalezienie się w ciemnym pokoju (i poddanie się zaskakującemu poczuciu głodu i deprywacji sensorycznej) jest bardzo zaskakującym stanem, mimo że reprezentuje środowisko z maksymalnie przewidywalnym wkładem sensorycznym. Z tego też powodu problem ciemnego pokoju jest fikcyjny. Co więcej, podkreślają Schwartenbeck z Fristonem, zgodnie z FEP podmioty wcale nie chcą szukać pozbawionych bodźców nisz, a raczej aktywnie eksplorować środowisko, ponieważ FEP traktuje celowe działania jako aktywne wnioskowanie. Pojęciem tym szerzej zajmę się w dalszej części tego rozdziału. Teraz należy tylko stwierdzić, że aktywne wnioskowanie minimalizuje zaskoczenie na drodze selektywnego próbkowania danych sensorycznych w taki sposób, aby można je było dostosować do generowanych predykcji. Oznacza to, że wcześniejsze przekonania (*priors*) podmiotu obejmują założenie, zgodnie z którym podejmie on odpowiednie działania minimalizujące zaskoczenie. Podmiot zatem musi reprezentować się w przyszłych stanach, wykonując określone działania (Friston i in., 2014; Schwartenbeck i in., 2013, s. 2). Ujęcie to wciąż koncentruje się jednak tylko na stronie podmiotu, nie wyjaśniając zupełnie, dlaczego środowisko jest dla niego znaczące.

Należy stwierdzić, że problem ciemnego pokoju budzi obawy nie tylko u krytyków PP, ale także wśród jego zwolenników, szczególnie tych, którzy wiążą je z eksplanacyjną funkcją FEP. Kiverstein, Miller i Rietveld (2019) zgadzają się z tym, że FEP oferuje dobre wyjaśnienie tego, dlaczego podmioty minimalizujące swoją energię swobodną unikają ciemnego pokoju, niemniej dostrzegają potrzebę takiej wykładni tej zasady, jaka pozwoli objaśnić możliwość takich typowych dla żywych stworzeń zjawisk jak ciekawość, zabawa czy poszukiwanie nowości. Innymi słowy, dlaczego podmiot jest zmotywowany, aby wybierać aktywną eksplorację, a nie bezpieczną i łatwo przewidywalną eksploatację? Friston w wielu pracach pokazuje, że w aktywnym wnioskowaniu podmiot nie tylko działa, aby utrzymać siebie w stanach, których oczekuje, ale także, by minimalizować niepewność związaną ze stanami przyszłymi (por. Schwartenbeck i in., 2013; Friston i in., 2014; Friston i in., 2017a). Możliwość balansu pomiędzy tymi działaniami ściśle wiąże się z czymś, co w ramach wnioskowania aktywnego określa się jako „zasady wyboru/zasady działania" (*policy selection*/*action selection*). Przez zasady działania rozumie się tutaj reguły wyboru działań i kolejności ich podejmowania w odniesieniu do oczekiwanych stanów, w jakich podmiot ma się znaleźć. Owe zasady mają służyć minimalizacji przyszłej energii swobodnej, przez którą

Friston i jego współpracownicy rozumieją oczekiwaną energię swobodną (*expected free energy*) (por. Friston i in., 2015b).

Po tym omówieniu przejdę teraz do krytyki PP i FEP przeprowadzonej przez Colina Kleina (2018), ponieważ są w niej widoczne wszelkie słabości tych ujęć. W dalszej części tego rozdziału pokażę, że proponowana tutaj koncepcja PP oparta na modelu normatywności mechanizmów predykcyjnych jest odporna na zarzuty postawione przez Kleina.

5.2.2. Krytyka przetwarzania predykcyjnego z perspektywy problemu ciemnego pokoju

Interesującą krytykę PP oraz FEP przeprowadził Klein w artykule pod znaczącym tytułem *What do predictive coders want?* Jego zdaniem cały problem ciemnego pokoju można sprowadzić do pytania o motywację: „jak to jest, że stany motywacyjne takie jak pragnienia mogą istnieć w ramach PP?" (Klein, 2018, s. 3; por. także Sun, Firestone, 2020). Istotne jest to, że motywacji nie da się sprowadzić do przewidywania. Klein przedstawia przykład, który unaocznia problem motywacji. Wyobraźmy sobie, że siedzimy w ciemnym pokoju. Jest nam w tym położeniu dobrze i czujemy się bezpieczni. Co może nas zmusić, byśmy go opuścili? Na przykład poczucie głodu. Jednak przewidywanie głodu nie jest tym samym co bycie motywowanym przez głód. Przecież leżąc w łóżku w takim pomieszczeniu, możemy doskonale przewidzieć, że w pewnym momencie najdzie nas potrzeba zjedzenia, a także to, że nie realizując jej, będziemy coraz słabsi, mniej aktywni, aż w pewnym momencie umrzemy. Jest to łatwo przewidywalny stan i nie wiąże się z nim żadne zaskoczenie. Z całą pewnością umrzemy. A jednak jesteśmy skłonni działać w inny sposób. Głód nie jest przecież jedynym stanem, który może nas zmotywować raczej do działania niż do pozostawania w stanie spoczynku.

Klein uważa, że problem ciemnego pokoju domaga się wyjaśnienia, dlaczego i w jaki sposób działamy, minimalizując niepewność. Na niewiele zdaje się tłumaczenie Hohwy'ego, który twierdzi, że nie musimy minimalizować każdego błędu predykcyjnego, ale średni (*average*) błąd predykcyjny dany w dłuższych okresach, a na to nie pozwala nam przebywanie w ciemnym pokoju (Hohwy, 2013, s. 85, 175)[5]. Wyobraźmy sobie, powiada autor *The predictive mind*, że błędnie postrzegamy psy jako owce. W dłuższej perspektywie czasowej niemożność

[5] Z średnim błędem predykcyjnym, inaczej mówiąc – ważoną sumą błędów predykcyjnych (*weighted sum of prediction errors*), należy wiązać następującą kwestię: z powodu braku dostępu do przyczyn sygnału zmysłowego zaskoczenia nie daje się minimalizować w sposób bezpośredni, zatem model generatywny wykonuje swoje działanie na drodze generowania predykcji, które odnoszą się do uśrednionego zaskoczenia w perspektywie długoterminowej (Hohwy, 2013, s. 85).

rozróżnienia psów i owiec zwiększy błąd predykcyjny, czyli oddali nas od naszych oczekiwanych stanów. Analogicznie jest z ciemnym pokojem. Należy zgodzić się z Hohwym, niemniej to, co jest problematyczne, to brak wskazania zjawiska bądź mechanizmu, który wyjaśni naszą motywację do opuszczenia ciemnego pokoju. Klein przekonująco pokazuje, że stworzenie przebywające w ciemnym pokoju ma mniejsze błędy predykcyjne niż istoty żyjące w innych środowiskach, tak więc jego średni błąd jest dużo mniejszy niż tych istot. Powołanie się na pewność śmierci na niewiele się zdaje, bowiem w dłuższej perspektywie każdy z nas umrze. Pytanie zatem nie dotyczy tego, czy unikamy (bądź nie) momentu śmierci, ale dlaczego niektórzy z nas ten moment odwlekają (opuszczając ciemny pokój), a inni nie? Same przewidywania nie wystarczą, abyśmy podjęli określone działania adaptacyjne. Jest to zapewne powód, dla którego zdaniem Kleina teorie działania, np. z zakresu psychologii potocznej, nie ograniczają się do podania jednej racji czy przyczyny działania, ale powołują się na wiele postaw propozycjonalnych, takich jak przekonania, pragnienia czy motywacje (por. podrozdział *Mechanizmy predykcyjne jako przyczyny a subosobowy poziom wyjaśniania*)[6].

Poruszony przez Kleina problem motywacji bezpośrednio dotyczy kompromisu pomiędzy eksploracją a eksploatacją, a także problemu działania w warunkach zwiększonej niepewności oraz podwyższonego ryzyka. Jak wytłumaczyć takie „dziwne" z perspektywy FEP i PP zachowania jak wybór przedmiotu trudniej osiągalnego (np. w sytuacji, gdy zamiast zjedzenia jagód z krzaków, decyduję się na zerwanie jabłka z wysokiego drzewa wymagającego wspięcia się na nie), podjęcia działania, które może zakończyć się naszą śmiercią albo kalectwem (np. sporty ekstremalne albo strajk głodowy), lub właśnie zamknięcie się w ciemnym, pozbawionym bodźców pokoju, zamiast przebywać w normalnych warunkach dla życia. Twierdzę, że dopóki PP (czasem wiązane z FEP) nie odpowie na te pytania, to ostatecznie okaże się tylko interesującą i ekstrawagancką teorią, która nie potrafi wyjaśnić zjawisk, która obiecuje wyjaśnić, a w zamian opiera się na analizowaniu działań i percepcji w sztucznych, sterylnych warunkach. Już wiemy, że sama wielopoziomowa architektura modelu generatywnego nie wystarcza, by wyjaśnić te zjawiska. Nie wystarczają również mechanizmy predykcyjne, regulujące mechanizmy precyzji, zasady działania oparte na posiadanej wiedzy itd. Konieczne jest uwzględnienie tego, co w rozdziale *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego* określiłem jako „ograniczenia środowiskowe". Dopóki

[6] Van der Cruys, Friston i Clark (2020) twierdzą, że na zarzut niewystarczalności samych przewidywań (i pośrednio problem wyjaśnienia motywacji) można odpowiedzieć w ten sposób, że predykcje mają charakter „optymistyczny". Co to oznacza? Modele generatywne muszą być zorientowane w sposób optymistyczny (*optimistically biased*), ponieważ rozkład prawdopodobieństwa stanów oczekiwanych jest realizowany dopiero wtedy, gdy oddziałujemy na świat. Oznacza to, że predykcje są z jednej strony allostatycznie wbudowane w organizację danego organizmu, z drugiej zaś, że muszą być one możliwie realistycznie zorientowane na świat, ponieważ dotyczą działań własnych organizmów i ich przewidywanych konsekwencji.

się ich nie uwzględni, dopóty nasze wyjaśnienie przynajmniej niektórych zjawisk percepcyjnych i poznawczych pozostanie niepełne oraz niesatysfakcjonujące[7].

Problem motywacji jest zatem ściśle powiązany z tym, co świat prezentuje podmiotowi. Jest on bowiem nie tylko źródłem informacji oraz niepewności, ale także określonych wartości, potencjalnych zysków, otoczeniem, które może stanowić dla nas pewną nagrodę albo może się dla nas okazać niebezpieczne i przykre. Świat jest nie tylko źródłem pobudzeń, ale również przedmiotem naszego zainteresowania i pragnienia. Kwestię tę badacze, szczególnie ci powiązani z badaniami Fristona, dostrzegli już jakiś czas temu i zaproponowali wiele istotnych rozwiązań i pojęć, które mogą pomóc zrozumieć, dlaczego podmiot raczej wybiera zróżnicowanie świata niżeli nagrodzoną pewnością i bezpieczeństwem monotematyczność ciemnego pokoju. Tym zagadnieniem zajmę się w najbliższych rozważaniach[8]. Następnie odpowiem na pytanie, czy proponowane rozwiązania prowadzą do wyjaśnienia możliwości wyboru pomiędzy eksploracją a eksploatacją.

5.3. Działanie w warunkach niepewności

Poruszane przeze mnie zagadnienia bezpośrednio wiążą się z problematyką podejmowania decyzji. Dlaczego wybieramy takie, a nie inne działania? Co powoduje, że decydujemy się na jedne rzeczy, innych zaś unikamy? Mówiąc językiem PP, należy stwierdzić, że podmiot powinien podjąć taką decyzję, jaka zminimalizuje, w optymalny sposób, czyli w perspektywie długoterminowej, pojawiający się błąd predykcyjny. Minimalizacja tego błędu jest możliwa przy założeniu hierarchicznej struktury modelu generatywnego, w którym informacje zmysłowe obecne na niższych poziomach modelu są interpretowane przez znajdujące się na wyższych poziomach oczekiwania (predykcje). W PP dokonuje się wyboru pomiędzy konkurującymi

[7] Zdaniem Manuela Baltieriego i Christophera L. Buckleya (2019) problem ciemnego pokoju bierze się z przyjęcia przez wielu badaczy PP błędnego (w ich mniemaniu) poglądu, zgodnie z którym, aby procesy poznawcze pozwalały na skuteczne działanie, konieczne jest tworzenie możliwie dokładnych modeli świata. W takim ujęciu celem podmiotu jest dopasowanie odgórnych predykcji do oddolnie docierających wrażeń. Stąd im mniej tych wrażeń (jak w ciemnym pokoju), tym lepiej. W podejściach enaktywnych zorientowanych na działanie, a nie poznanie (np. te wykorzystujące model aktywnego wnioskowania), model generatywny nie musi kodować prawdziwych informacji o nadchodzących doznaniach, a wręcz przeciwnie, opisuje on pragnienia podmiotu, które mogą zostać osiągnięte przez aktywne wnioskowanie. Z perspektywy przedstawionej przeze mnie krytyki pytanie o to, dlaczego podmiot chce realizować swoje pragnienia, pozostaje otwarte. Inną kwestią jest problem tego, czy bez (chociaż niektórych) prawdziwych przekonań (w sensie bayesowskim) jest możliwe skuteczne działanie w środowisku? Ze względu na przyjętą w tej pracy perspektywę odpowiedź musi być negatywna (por. na ten temat podrozdział *Przetwarzanie predykcyjne a mechanistyczny model wyjaśniania naukowego*; a także Godfrey-Smith, 1996, rozdział 6).

[8] Pełna analiza odpowiedzi, jaką zwolennicy FEP formułują w odniesieniu do problemu motywacji oraz kompromisu eksploracji i eksploatacji, znacznie wykracza poza ramy niniejszych rozważań. Należy jednak zaznaczyć, że taka analiza jest trudna ze względu na nieustanne modyfikowanie i rozbudowywanie wyjściowych propozycji spod znaku tej ramy badawczej. Przedstawiona przeze mnie prezentacja jest zatem w znacznym stopniu ograniczona i należy ją traktować jako charakterystykę pewnych ogólnych tendencji obecnych w tym nurcie.

prognozami pochodzącymi z wyższego poziomu a dotyczącymi oczekiwanych stanów zmysłowych. Według Fristona (Friston i in., 2014) ów wybór jest procesem decyzyjnym, który należy rozumieć w kategoriach hierarchicznego aktywnego wnioskowania. Proces decyzyjny jednak nie jest prostym wyborem pomiędzy jedną a drugą predykcją, ale rozstrzygnięciem, czy należy wybrać krótkoterminową, czy długoterminową redukcję błędu predykcyjnego. Globalne, długoterminowe i ogólne predykcje mają tendencję do reagowania na błędy predykcyjne z dużym zyskiem. Natomiast bardziej lokalne, krótkoterminowe i detaliczne predykcje będą reagować na błędy przy mniejszym wzmocnieniu[9]. Duży zysk przemawia na korzyść uczenia się w sensie globalnym, gdzie model stara się objąć to, co dzieje się w stosunkowo szerokich, ale ubogich w szczegóły uogólnieniach, tzn. raczej rozpoznaje wzorce niż poszczególne szczegóły. Mniejszy zysk natomiast jest raczej ukierunkowany na zlokalizowane wzorce uczenia się, w których indywidualne cechy poszczególnych elementów skupiają się kosztem dopasowania ich do większych wzorców (Hohwy, 2013, s. 69; por. podrozdział *Rzeczywiste wzorce w przetwarzaniu predykcyjnym*).

Obraz ten jednak wydaje się niepełny w kontekście omawianego tutaj problemu ciemnego pokoju. Wybór pomiędzy krótko- a długoterminową redukcją błędu predykcyjnego albo jest optymalnym wyborem regulowanym zawsze i każdorazowo przez mechanizm minimalizacji niepewności, co oznacza, że zawsze wydajniejsze jest minimalizowanie błędów predykcyjnych w perspektywie długoterminowej (niezależnie od tego, czy podmiot dobrze dopasuje odpowiednią predykcję do uśrednionego błędu predykcyjnego), albo faktycznym wyborem pomiędzy redukcją krótko- a długoterminową, która jest zależna od preferencji i przekonań podmiotu. Mam na myśli to, że czasem możemy chcieć wybrać taką predykcję lub działanie, jakie będzie prowadziło do krótkoterminowej redukcji niepewności. Innymi słowy, może być tak, że z pewnego powodu raczej wybierzemy spoczynek w ciemnym pokoju aniżeli przebywanie w normalnym środowisku. Sam mechanizm predykcyjny modelu generatywnego nie wystarczy, by wyjaśnić, dlaczego pomimo względnie obiektywnego zysku w wyborze pewnych działań bądź predykcji podmiot może wybrać takie działania, jakie nie prowadzą do optymalnego zmniejszenia niepewności i zaskoczenia. Dlaczego czasem jesteśmy skłonni podjąć głodówkę albo wybrać pożywienie, którego zdobycie naraża nas na niebezpieczeństwo? Wielu badaczy już rozpoznało te trudności i z tego powodu uznali, że należy wskazać jeszcze inne elementy konstytutywne dla wyjaśnienia możliwości decydowania i działania w ramach PP. Omówię teraz te propozycje.

[9] W tym kontekście błąd predykcyjny jest rozumiany jako sygnał uczenia się, w którym wzmocnienie moduluje przebieg uczenia.

5.3.1. Wnioskowanie aktywne

Większość badań w PP nad procesami decyzyjnymi wiąże się z dociekaniami autorów, którzy w jakimś stopniu uznają zasadność FEP i ramy aktywnego wnioskowania. To ostatnie pojęcie jest dość niejasne, ponieważ Friston w różnych artykułach modyfikował je i nadawał mu kolejne znaczenia i funkcje. W pracy *The free-energy principle: A unified brain theory?* (2010, s. 129) Friston powiada, że potencjalne błędy predykcyjne (które w tym ujęciu można sprowadzić do wariacyjnej energii swobodnej) można minimalizować na dwa sposoby: (1) na drodze wnioskowania pasywnego (percepcyjnego), czyli rewizji modelu generatywnego (zmiana jego parametrów) oraz sformułowanych hipotez w odniesieniu do statystycznego sygnału zmysłowego; (2) na drodze wnioskowania aktywnego, czyli poprzez takie działanie w świecie, jakie pozwoli utrzymać odpowiednią, sformułowaną przez model hipotezę, w taki sposób, aby „nasze predykcje stały się prawdziwe" (Clark, 2016, s. 121). Chodzi o taką zmianę w środowisku, że zmieni się struktura przyczynowa określonych stanów świata (poprzez selektywne próbkowanie niespodziewanych wrażeń) (por. Badcock, Friston, Ramstead, 2019, s. 8; Friston, Daunizeau, Kiebel, 2009)[10]. Znaczy to, że wykonując wnioskowanie aktywne, podmiot ingeruje w przyczynową strukturę odpowiednich stanów rzeczy. Wnioskowanie aktywne jest rozumiane, w pewnym sensie, jako takie działanie, jakie minimalizuje niepewność. Tego typu działania są adaptacyjne, ponieważ ich celem jest (w zgodzie z FEP) utrzymanie homeostazy. Z tego powodu można powiedzieć, że wnioskowanie aktywne jest zastosowaniem FEP do działania. Polega ono na tym, że podmiot chce pozostać przy życiu, utrzymując swoją homeostazę[11]. W tym celu musi upewnić się, że takie ważne parametry, jak temperatura ciała lub natlenienie krwi, nie odbiegają zbytnio od normy, tj. nie są zaskakujące. Parametry te jednak mogą być tylko wywnioskowane z pomiarów sensorycznych, podmiot zatem jest zmuszony oddziaływać na środowisko, zmniejszając tym samym niepewność. Innymi słowy, działanie może bezpośrednio wpływać na świat, zatem różne działania mogą regulować różne formy przyszłości i wpływać na kształt sygnału zmysłowego i powiązanych z nim możliwych błędów predykcyjnych. Takie ujęcie aktywnego wnioskowania określam jako „wąskie" i można go używać zamiennie z pojęciem działania, rozumiejąc przez to działania, które minimalizują błędy predykcyjne.

W nowszych ujęciach wnioskowanie aktywne wiązane jest z kocami Markowa i wariacyjną neuroetologią (Ramstead i in., 2017; Ramstead i in., 2018). Takie podejście można określić jako „szerokie aktywne wnioskowanie" bądź jako „rama

[10] Należy dodać, że wnioskowanie percepcyjne określa ograniczenia (poprzez modyfikację odpowiednich parametrów modelu) dla wnioskowania aktywnego (Hohwy, 2020a, s. 8).

[11] Oznacza to, że jeżeli FEP jest zasadą normatywną dla wszelkich żywych organizmów, to aktywne wnioskowanie jest ramą normatywną wyjaśniającą optymalność działań i zachowań tych organizmów (por. Schwartenbeck i in., 2013).

aktywnego wnioskowania". Przypomnijmy: każdy (ergodyczny) system, który istnieje, musi, z racji swojego istnienia, być otoczony utrzymującym go przy istnieniu kocem Markowa. Wszelkie systemy biologiczne można opisać jako koce Markowa mające stany wewnętrzne, które same składają się z kolejnych koców Markowa i ich stanów wewnętrznych itd. Ten formalizm można powtórzyć w górę i w dół, w odniesieniu do różnorodnych, zagnieżdżonych skal organizacji, w których istnieją systemy biologiczne. W ten sposób systemy biotyczne (tj. pojedyncze komórki, organizmy, grupy społeczne i kulturowe) można opisać jako (wielowymiarową) przestrzeń fazową indukowaną przez hierarchię koców Markowa. Dobrym przykładem są ludzie, których mózgi, narządy zmysłowe, mięśnie, komórki itd. są kocami Markowa z własnymi granicami (Hesp i in., 2019, s. 2001; por. także Kirchhoff i in., 2018). Podejście to jest powiązane z FEP i tutaj przez aktywne wnioskowanie rozumie się nie tyle działania, ile całą ramę badawczą, która jest wiązana z dynamicznym modelowaniem przyczynowym (*dynamic causal modelling*, por. Friston, Harrison, Penny, 2003)[12]. Rama aktywnego wnioskowania ma dostarczyć normatywne wyjaśnienie, dlaczego agenci muszą z konieczności wnioskować i minimalizować zaskoczenie, wykorzystując swoje wewnętrzne, ukryte stany, w celu zachowania integralności fizjologicznej (Friston, 2012; Morville i in., 2018, s. 16). Trzeba dodać, że formalizm aktywnego wnioskowania ma zapewniać integracyjne ramy dla wielu normatywnych podejść (Friston i in., 2017a, s. 4)[13].

Zwolennicy FEP[14] proponują, aby procesy percepcyjne, decyzyjne, uczenia się czy kontroli motorycznej analizować w kategoriach hierarchicznego aktywnego wnioskowania (por. Friston, Daunizeau, Kiebel, 2009; Friston, Mattout, Kilner, 2011; Friston i in., 2012b; Friston i in., 2015b; Friston i in., 2016; Friston i in., 2017a; Friston i in., 2017b; Schwartenbeck i in., 2013). Friston opisuje decyzje jako przekonania na temat alternatywnych zasad (polityk) działania. Zasady działania są rozumiane również jako zasady epistemiczne, ponieważ decydują one o wyborze sekwencji działań (Pezzulo i in., 2016, s. 323). To, które działanie lub ich sekwencja zostanie wybrane, zależne jest od oczekiwań dotyczących sygnału zmysłowego. Zasady są wybierane przy wcześniejszym przekonaniu, że minimalizują

[12] W niniejszych analizach nie będę zbyt wchodził w zawiłości powiązania aktywnego wnioskowania z wariacyjną energią swobodną i neuroetologią oraz etologią wariacyjną. Badania te znacznie wykraczają poza ramy przeznaczone dla tej pracy. Ograniczę się zatem do takiego rozumienia ramy aktywnego wnioskowania, zgodnie z którym jest ono ramą normatywną i eksplanacyjną, która zakłada, że podmiot działa optymalnie w odniesieniu do swojego modelu generatywnego wtedy, gdy wybiera taką sekwencję działań, które minimalizują oczekiwaną energię swobodną, *resp.* błąd predykcyjny, tj. minimalizują oczekiwane zaskoczenie (niepewność) lub (równoważnie) maksymalizują ewidencje modelu (Constant i in., 2019, s. 14).

[13] Friston i współautorzy (2017a) przedstawili schemat aktywnego wnioskowania dla dyskretnych modeli wyboru przestrzeni w przestrzeni stanów, który jest odpowiedni do modelowania różnych paradygmatów i zjawisk. Oparta na nim teoria procesu ma zapewniać prosty opis wielu zjawisk empirycznych.

[14] W niniejszych analizach rekonstruuję rozumienie aktywnego wnioskowania jako ramy, która jest w jakimś stopniu komplementarna z PP (por. m.in. Bruineberg, Kiverstein, Rietveld, 2016; Friston, Kiebel, 2009; Hohwy, 2013, 2016, 2020). Nie analizuję tutaj stanowisk krytycznych wobec tego powiązania (por. m.in. Colombo, Wright, 2018; Klein, 2018) bądź nastawionych agnostycznie (por. m.in. Clark, 2016; Orlandi, 2016; Sims, 2017).

one błędy predykcyjne pomiędzy osiągalnymi i pożądanymi wynikami na wielu zagnieżdżonych hierarchicznie poziomach. Zasady działania (*policy selection*) nie opierają się tylko na wymogu aktualizacji przekonań bayesowskich, ale także na konieczności podjęcia określonych działań. Oznacza to, że wybór zasad działania zależy od: (1) potencjału zdobywania informacji o przyszłych stanach świata (tj. wartość epistemiczna – „gdzie powinienem być, jeżeli to a to"), oraz (2) potencjał osiągania preferowanych rezultatów zmysłowych (tj. wartość pragmatyczna – „co powinienem postrzegać, jeżeli to a to") (Constant i in., 2019). Co ważne, zasady są wybierane na podstawie przekonania zarówno o ich oczekiwanym wyniku, jak i oczekiwanej precyzji (Burr, 2017, s. 9). Należy stwierdzić, że im bardziej zminimalizowana zostaje energia swobodna, *resp.* błąd predykcyjny, tym prawdopodobniejsze i skuteczniejsze okazują się określone zasady (polityka) działań. Energia swobodna ważona jest przez precyzję, która określa, jak bardzo jesteśmy pewni naszych przekonań dotyczących zasad działania. Jest to istotne, kiedy chcemy określić, czy wybieramy eksplorację czy eksploatację, ponieważ im większą mamy pewność w odniesieniu do zastosowania naszych zasad (tzn. jak dużą precyzję się im przypisze), tym mniej eksplorujemy, a więcej eksploatujemy. Im mniejszą precyzję przypiszemy naszym zasadom, tym więcej szukamy nowości i jesteśmy skłonni zwiększyć poziom niepewności oraz zaskoczenia (Solopchuk, 2018).

Docieramy tutaj do kluczowej kwestii poruszanej w niniejszym rozdziale, czyli kompromisu pomiędzy eksploatacją a eksploracją. Odpowiedź na ten problem jest zależna, co już jasno widać, od tego, jakie zasady działania wybierzemy. Te zaś nie są tylko zależne od (bayesowskiej) sieci naszych przekonań, ale także od pewnych wartości, które oferuje środowisko. Temu problemowi przyjrzę się bliżej.

5.3.2. Podejmowanie decyzji

Friston i jego zwolennicy wielokrotnie podkreślają, że ich koncepcja zrywa z klasycznymi, normatywnymi ujęciami, opartymi na pojęciach optymalnego działania, użyteczności i oczekiwanej wartości. Proponowane ujęcie opiera się na bayesowskim (aktywnym) wnioskowaniu i maksymalizacji wiarygodności modelu[15]. Aby dobrze zrozumieć tę propozycję, należy krótko scharakteryzować tzw. podejście klasyczne. Friston bowiem i jego współpracownicy nieustannie konfrontują swoją koncepcję z ujęciami klasycznymi. Drugim celem ataku zwo-

[15] Można by w tym miejscu angielskie *evidence* tłumaczyć jego polską kalką, jaką jest ewidencja, ponieważ polskie słownikowe odpowiedniki (tj. dowód, świadectwo) nie oddają dobrze sensu tego terminu. Model jest ewidentny, kiedy jest on wierzytelny (w sensie pewnej „wierności" danym i bycia oczywistym dla siebie), lepiej szacujący prawdopodobieństwa, bardziej dopasowany do odpowiednich zmiennych. *Wiarygodność* bądź *wierzytelność* wydają się zatem odpowiednimi przekładami *evidence*, zachowują bowiem jego semantyczną wieloznaczność.

lenników FEP i aktywnego wnioskowania są koncepcje uczenia nadzorowanego. Zajmę się nimi niebawem.

5.3.2.1. Klasyczne modele

W klasycznej ekonomii procesy podejmowania decyzji w warunkach niepewności są modelowane matematycznie, wykorzystując teorię oczekiwanej wartości i teorię oczekiwanej użyteczności (Fishburn, 1988; Tyszka, 2010, s. 191–222). Zgodnie z pierwszą z nich oczekiwana wartość jest sumą wszystkich możliwych do uzyskania wpłat (tj. zysków i strat) przemnożonych przez prawdopodobieństwo ich uzyskania. Z tej perspektywy zasada maksymalizacji oczekiwanej wartości zaleca wybór takiego działania, jakie ma największą wartość oczekiwaną. Jest to koncepcja, którą wiąże się z obiektywną miarą wartości konsekwencji[16].

Już w XVIII wieku Daniel Bernoulli doszedł do wniosku, że ludzie w swoich decyzjach nie stosują zasady maksymalizacji oczekiwanej wartości. Zastąpił ją pojęciem oczekiwanej użyteczności, którą określa się jako subiektywną miarę wartości. Z teorią tą wiąże się hipoteza, zgodnie z którą ludzie mają lub zachowują się w taki sposób, jakby mieli funkcję użyteczności. Jest ona definiowana na zbiorze pewnych alternatyw i w sytuacji ryzyka. Zgodnie z tym ujęciem, kiedy trzeba wybrać jakieś losowe zdarzenie z wynikami w tym zbiorze, ludzie zachowują się tak, jakby chcieli zmaksymalizować oczekiwaną wartość tej funkcji. W praktyce wygląda to tak, że każdemu potencjalnemu elementowi zbioru możliwych wyborów podmiot przyporządkowuje pewną wartość użyteczności i na podstawie tej wartości podejmuje decyzje.

John von Neumann i Oskar Morgenstern (1944) zaproponowali, aby określić użyteczność na podstawie porównywania loterii, czyli takich alternatyw, które zawierają nie tylko określone konsekwencje, ale także prawdopodobieństwa tych konsekwencji. Loteria formalnie oznacza dowolny rozkład prawdopodobieństwa na znanym i skończonym zbiorze wyników. Takimi wynikami mogą być dowolne konsekwencje wyborów, np. jakieś sumy pieniędzy, nagrody czy zdarzenia. Możliwość porównywania loterii pozwala określić użyteczność. To jednak wymaga od decydenta spełnienia określonych warunków racjonalności, które określa się mianem „aksjomatów". I tak relacja preferencji danego decydenta $x_i \leq x_j$ określona na zbiorze *P*, którego elementami są takie zmienne losowe, musi spełniać cztery aksjomaty (por. Tyszka, 2010, s. 197–198):

[16] Zasada ta jednak nie musi przesądzać o danym wyborze. Jej słabości wiążą się z tzw. paradoksem petersburskim, który wykazuje, że przy podejmowaniu decyzji liczy się nie tylko zyski jako takie (np. bogactwo), ale także użyteczność, np. pożytek, przyjemność czy zadowolenie.

1. Aksjomat zupełności (kompletności) mówi, że relacja preferencji gracza jest spójna. Oznacza to, że decydent potrafi porównać ze sobą dwie loterie, które ma do wyboru. Tj. dla każdej pary alternatyw x_i, x_j (w tym loterii) decydent albo słabiej preferuje x_i niż x_j ($x_i \geq x_j$), albo słabiej preferuje x_j niż x_i ($x_j \geq x_i$), albo zachowuje obie relacje jednocześnie.

2. Aksjomat przechodniości mówi, że relacja preferencji decydenta jest przechodnia. Tj. dla każdego x_i, x_i, x_k, jeśli $x_i \geq x_j$ oraz $x_j \geq x_k$, to musi też zachodzić $x_i \geq x_k$.

3. Aksjomat ciągłości mówi, że relacja preferencji decydenta jest ciągła w taki sposób, że dla każdej trójki wyników takich, że $x_i \leq x_j \leq x_k$ istnieje takie prawdopodobieństwo p, że x_j jest w tym samym stopniu preferowane co loteria (x_i, p; x_k, 1-p).

4. Aksjomat niezależności mówi, że jeżeli jeden decydent jest indyferentny pomiędzy jakimiś dwoma wynikami x_i i x_j, to powinien być również indyferentny pomiędzy loteriami, które dają x_i i x_j z tym samym prawdopodobieństwem.

$$\text{Jeżeli } x_i \sim x_j, \text{ to } px_i + (1\text{-}p)\, x_k \sim px_j + (1\text{-}p)x_k \text{[17]}.$$

Zgodnie z teorią oczekiwanej użyteczności, jeżeli preferencje decydenta spełniają powyższe aksjomaty, to można je reprezentować za pomocą funkcji użyteczności $u()$ zdefiniowanej na zbiorze wyników x_i lub rozszerzonej na zbiór loterii $L=(x_i,p_i; \ldots; x_n,p_n)$, gdzie $\sum_{i=1}^{n} p = 1$. Rozszerzona funkcja $U()$ reprezentuje preferencje na loteriach w taki sposób, że L1≥L2 wtedy i tylko wtedy, gdy $U(L1) \geq U(L2)$, przyjmując postać użyteczności oczekiwanej (tj. liniowej względem prawdopodobieństw): $U(L1) = EU_L = \sum_{i=1}^{n} p_i u(x_i)$. Wzór ten mówi tyle, że maksymalizacja oczekiwanej funkcji użyteczności prowadzi do wyboru najbardziej preferowanej alternatywy.

Aksjomaty teorii oczekiwanej użyteczności oparte są na założeniu, że podmioty są racjonalne i mają dobrze zdefiniowane preferencje. Wiąże się to z faktem, że akceptowalny poziom ryzyka bądź niepewności jest ściśle zależny od indywidualnego stosunku do tych zmiennych.

Teorie oczekiwanej wartości oraz oczekiwanej użyteczności określa się mianem „normatywnych", bowiem ich celem jest pokazanie, w jaki sposób ludzie powinni zachowywać się, kiedy muszą podjąć decyzję w warunkach niepewności bądź ryzyka. Dobrym przykładem takiego ujęcia jest model subiektywnie oczekiwanej użyteczności Leonarda J. Savage'a (1954), określany również jako bayesowski model podejmowania decyzji. Zdaniem tego badacza podmiot wybiera pomiędzy alternatywnymi działaniami, których konsekwencje są zależne od tego, jaki z wykluczających się stanów rzeczy zostanie zrealizowany[18]. Trudność tego modelu

[17] Odpowiednikiem tego aksjomatu jest w decyzjach w warunkach niepewności tzw. zasada rzeczy pewnej (*sure thing principle*), zgodnie z którą, jeżeli przy pewnym stanie świata otrzymamy taki sam wynik, niezależnie od tego, co się wybierze, to nasz wybór nie powinien zależeć od tego wyniku (Tyszka, 2010, s. 199).

[18] Chodzi o zbiór stanów rzeczy, które się wzajemnie wykluczają i jednocześnie wykluczają wszystkie możliwości.

polega na tym, że podmiot musi wybierać w warunkach niepewności, bowiem nie da się podać obiektywnego prawdopodobieństwa zrealizowania się odpowiednich stanów rzeczy. Podmiot, nie znając obiektywnego prawdopodobieństwa, musi podjąć jakąś decyzję. Zgodnie z omawianym modelem, twierdzi się, że ma on określone preferencje w doborze działań. Podstawową z nich ma być dążenie do maksymalizacji subiektywnie oczekiwanej użyteczności. Podmiot zatem ma wybierać takie działanie, jakiego subiektywnie oczekiwana użyteczność jest największa. W tej perspektywie można stwierdzić, że podmiot jest racjonalny, kiedy działa zgodnie z tak rozumianą preferencją[19].

Badacze podkreślają, że problem z tego typu teoriami normatywnymi polega na tym, że czym innym jest stwierdzenie, że ludzie w takich, a takich warunkach powinni podejmować takie, a nie inne decyzje, a czym innym powiedzenie, w jaki sposób ludzie się naprawdę zachowują i jakie decyzje faktycznie podejmują[20]. Z tego też powodu zaproponowano teorie deskryptywne, które empirycznie i eksperymentalnie dowodzą, w jaki sposób ludzie się zachowują w różnych sytuacjach. Przykładem takiej koncepcji jest teoria perspektywy Daniela Kahnemana i Amosa Tversky'ego (1979, 1992). Badacze ci wskazali na pewne słabości teorii normatywnych, podkreślając, że nie jest tak, że ludzie zawsze próbują maksymalizować użyteczność swoich działań (por. podrozdział *Uwagi wobec ramy aktywnego wnioskowania*). Wiąże się to m.in. z tzw. efektem pewności, zgodnie z którym ludzie przypisują większe znaczenie takim działaniom, jakich wyniki są pewne, a nie mniej lub bardziej prawdopodobne.

5.3.2.2. Krytyka klasycznych modeli z perspektywy koncepcji wnioskowania aktywnego

Należy stwierdzić, że krytyka Fristona i jego zwolenników wymierzona jest w teorie normatywne w podanym tutaj znaczeniu. Nie wchodzi jednak w dyskusję z teoriami deskryptywnymi, co jak wykażę, ma określone negatywne konsekwencje. Na początku trzeba również zwrócić uwagę na fakt, że odrzucenie przez zwolenników FEP klasycznych modeli podejmowania decyzji jest przeprowadzone w sposób pobieżny i ogólny. Nie tyle wykazuje się ich słabość eksplanacyjną bądź teoretyczną, ile odrzuca się je jako mniej satysfakcjonujące niż normatywny model oparty na FEP. Można stwierdzić, że nie tyle sama konstrukcja teorii normatywnych

[19] Na temat pojęcia racjonalności zob. podrozdział *Normatywność bayesowskiego modelu przetwarzania predykcyjnego*.

[20] Wiele badań empirycznych wykazało, że rzeczywiste decyzje ekonomiczne bardzo często nie są zgodne z aksjomatami wskazanymi przez von Neumanna i Morgensterna, przez co mogą pojawić się tzw. paradoksy racjonalności. Ich przykładem są paradoks Allaisego, który dotyczy zbytniego preferowania przez ludzi pewnego zysku; albo paradoks Ellsberga, zgodnie z którym ludzie nadmiernie unikają wyników niepewnych. Por. także przypis 17, s. 20, dotyczący równowagi Nasha.

jest nieprzekonująca, ile leżąca u ich podstaw zasada (w różnych wersjach różnie sformułowana i różnie akcentowana) zakładająca istnienie takiej funkcji wartości, jaka musi być maksymalizowana, aby mogła wyjaśnić zachowania i działania.

W ramach FEP twierdzi się, że warunkiem wyjaśnienia ludzkich działań[21] oraz wielu innych zjawisk, takich jak wyższe procesy decyzyjne, relacje międzyosobowe, kształtowanie ja oraz poznanie społeczne (por. Constant i in., 2019; Gallagher, Allen, 2016; Hesp i in., 2019; Khalvati i in., 2019; Moutoussis i in., 2014a; Moutoussis i in., 2014b) jest odniesienie się do pojęcia wartości, która jest tutaj rozumiana jako dowód dla modelu, czyli jako rozkład brzegowy funkcji wiarygodności (Friston, Adams, Montague, 2012). Ten zaś jest produktem wiarygodności dla pojedynczych obserwacji. W tej funkcji pewne zmienne zostają zmarginalizowane. Funkcja ta pozwala na (1) szacowanie parametrów, które mają największą wiarygodność (w przypadku PP będą to predykcje i przekonania o wysokim prawdopodobieństwie, czyli wiarygodniejsze); oraz (2) na testowanie hipotez statystycznych (czyli zwiększanie wierzytelności modelu na drodze aktywnego wnioskowania). Dla modelu wartością jest zatem to, co maksymalizuje jego wiarygodność. Wystarczającym wyjaśnieniem zachowania bądź działania jest gromadzenie dowodów, które zwiększają wiarygodność modelu. W ten sposób działania można wyjaśnić przy pomocy pojęcia wnioskowania aktywnego: podmioty aktywnie testują środowisko, aby w ten sposób zwiększyć wiarygodność swoich modeli świata. Sposób, w jaki to robią, wyznaczają odpowiednie zasady działania, które są podporządkowane imperatywowi minimalizacji niepewności.

W tym ujęciu każde działanie rozumiane jest jako bayesowskie aktywne wnioskowanie ograniczane przez wcześniejsze przekonania. Te zaś są nieustannie optymalizowane w odniesieniu do hierarchii poziomów modelu. Fakt ten jest uzasadniany przez pełne twierdzenie klasowe (*complete class theorem*), zgodnie z którym każdy możliwy zestaw zachowań lub działań można wyjaśnić w odniesieniu do (przynajmniej jednego) zbioru przekonań (*priors*) (Friston, Adams, Montague, 2012, s. 6). Za każdym razem możemy wskazać model bayesowski, zgodnie z którym dane zachowanie bądź działanie wydaje się optymalne. Ten rodzaj optymalności jest zależny od modelu, tzn. jest subiektywny, ale nie jest arbitralny, bowiem wiąże się z fenotypem danego organizmu. W praktyce oznacza to, że aktywne wnioskowanie nie zakłada optymalizacji w żadnym absolutnym sensie. Takie rozwiązanie ma pozwalać na uniknięcie tautologii wpisanej w klasyczne modele normatywne[22]. Zdaniem Fristona i współpracowników tautologia ta polega na tym, że zgodnie

[21] W ujęciu bayesowskim optymalne zachowanie bądź działanie danego organizmu jest rozwiązaniem określonego wnioskowania probalistycznego, które dotyczy zadania wynikającego z jego kontaktu ze środowiskiem. W ramie aktywnego wnioskowania optymalne działania można opisywać w terminach predykcji ukierunkowanych na oczekiwane konsekwencje (Friston i in., 2010, s. 235). Należy jednak podkreślić, że coś jest optymalne w tym sensie, że konstrukcja wariacyjnej energii swobodnej sprowadza problem wnioskowania do problemu optymalizacji.

[22] Takie jak teorii oczekiwanej wartości i teorii oczekiwanej użyteczności.

z tymi modelami, jeżeli zachowanie jest optymalne, to maksymalizuje wartość. Wartość zaś jest tym, co czyni zachowanie optymalnym. W takim ujęciu trudno jednak wyjaśnić pochodzenie wartości jako takiej inaczej niż poprzez odniesienie się do parametrów w danym modelu. W modelu podejmowania decyzji opartym na aktywnym wnioskowaniu problem ten rozwiązano przez odwołanie się do hierarchicznego wnioskowania bayesowskiego, w którym optymalizacja na jednym poziomie jest ograniczona przez empiryczne przekonania (*empirical priors*) z wyższego poziomu. Optymalizacja (wartość) odnosi się zatem do maksymalizacji dowodów modelu bayesowskiego lub (jednocześnie) do minimalizacji zaskoczenia. Optymalne działanie (w sensie podanym wyżej) to takie, jakie na podstawie wcześniejszych przekonań pociąga za sobą przewidywanie przyszłych konsekwencji w celu dostępu do określonych wartości (Friston, Adams, Montague, 2012, s. 2). W ten sposób maksymalizowana jest wiarygodność modelu, którą należy rozumieć jako oczekiwaną nagrodę dla przyszłych stanów podmiotu.

W tym ujęciu aktywne wnioskowanie wiąże się z minimalizacją energii swobodnej i pozwala na powiązanie teorii optymalnego działania z FEP. Działania oparte na przekonaniach podmiotu minimalizują energię swobodną, która wiąże się z marginalnym prawdopodobieństwem obserwowanych stanów rzeczy. W ten sposób maksymalizują one owo prawdopodobieństwo lub wiarygodność modelu.

Rama aktywnego wnioskowania wymaga sformułowania definicji minimalizacji energii swobodnej w czasie dyskretnym i zastanowienia się nad tym, w jaki sposób nagroda lub funkcje kosztowe są uwzględniane[23]. Powiązanie niniejszej koncepcji z FEP pozwala na sformułowanie normatywnej teorii działania i percepcji. Jest ona normatywna, ponieważ zapewnia funkcję celową (tj. wariacyjną energię swobodną), względem której model się optymalizuje zarówno pod względem działania, jak i percepcji[24]. Wydaje się jednak, że ostatecznie źródło normatywności w FEP sięga jeszcze głębiej. Mianowicie: nie tyle chodzi o samą tę funkcję, ile o jej status jako warunku możliwości istnienia samoorganizujących się systemów. Model zatem nie tylko faktycznie się optymalizuje względem tej funkcji, ile *musi* się wobec niej optymalizować. Inaczej nie będzie istniał (por. Colombo, Wright, 2018; Hohwy, 2020b).

W ujęciu FEP, które ma wskazać biologiczno-informacyjną podbudowę dla teorii działania, optymalne decyzje minimalizują różnicę pomiędzy prawdopodobnymi a oczekiwanymi stanami systemu. Oznacza to, że czasami podmioty są zmuszone szukać nowych doznań, zaś w innych sytuacjach maksymalizują oczekiwaną użyteczność. Aby podjąć decyzję, podmiot musi działać zgodnie z FEP, co w praktyce oznacza, że eksploracja polega na minimalizacji zaskoczenia bądź maksymalizacji

[23] Z tego też powodu zwolennicy omawianego ujęcia postulują powiązanie badań nad aktywnym wnioskowaniem z analizami koców Markowa. Ze względu na ograniczenia nałożone na tę pracę niniejszy wątek w zasadzie pomijam.

[24] Owa normatywna funkcja ma być wzmocniona (według omawianych autorów) przez neuronalnie wiarygodny schemat PP (Schwartenbeck i in., 2013).

wiarygodności modelu (Schwartenbeck i in., 2013). Dla tych badaczy oznacza to, że wybór zasad działania przez podmiot (który nie jest pewien konsekwencji swoich działań) wiąże się z konieczności z potrzebą eksploracji, która towarzyszy tradycyjnie rozumianej maksymalizacji oczekiwanej użyteczności. Innymi słowy, wymóg minimalizacji niepewności, *resp.* energii swobodnej, implikuje zachowania eksploracyjne, czyli takie, jakie poszukują wartości (ta zaś powoduje pożądaną i homeostatycznie warunkowaną maksymalizację wiarygodności modelu).

W ten sposób FEP radzić ma sobie z problemem ciemnego pokoju oraz kompromisem eksploracji i eksploatacji: podmiot nie chce przebywać w ciemnym pokoju, ponieważ takie położenie jako zaskakujące nie pozwala mu na zwiększenie wiarygodności własnych przekonań. Albo zatem zapali światło, tym samym zamieniając zaskakujący stan rzeczy (ciemny pokój) na oczekiwany, albo opuści niepożądane przez niego pomieszczenie. W obu tych przypadkach działanie (aktywne wnioskowanie) zwiększa wiarygodność modelu, tym samym minimalizując niepewność. Zapalenie światła albo opuszczenie ciemni ma wartość, która pozwala o takich działaniach orzekać, że są one optymalne.

W tym miejscu ujawnia się główna różnica pomiędzy normatywną propozycją opartą na FEP i wnioskowaniu aktywnym a klasycznymi modelami podejmowania decyzji konstruowanymi wokół pojęcia oczekiwanej użyteczności. W tych modelach decyzja jest optymalna tylko wtedy, gdy maksymalizuje użyteczność. W modelu opartym na FEP decyzja jest optymalna (w sensie względnym, tj. zgodnie z pełnym twierdzeniem klasowym) wtedy, gdy zasady ją określające (*policy*) są wybierane w celu zminimalizowania różnicy pomiędzy oczekiwanymi konsekwencjami działań a rozkładem prawdopodobieństwa określonych zmiennych w środowisku (Schwartenbeck i in., 2013, s. 3). Należy pamiętać, że podmiot musi ustalić, które zasady najprawdopodobniej będzie realizować, a następnie wybrać ich odpowiedni zbiór. Problem ten wiąże się z tym, w jaki sposób podmiot się uczy, czyli jak jego wewnętrzny model zmienia swoje parametry (poprzez modyfikację starych przekonań bądź nabywanie nowych).

Omawiane ujęcie pozwala również wyjaśnić pojęcie nagrody, kluczowe dla zrozumienia możliwości aktywnego poszukiwania przez podmiot nowości bądź unikania nudy. Nagrodę zewnętrzną odnosi się do jakiegoś zewnętrznego względem podmiotu wzmocnienia. Odpowiada ona za maksymalizację oczekiwanej użyteczności (eksploatacja). Nagroda wewnętrzna zaś odnosi się do maksymalizacji wiarygodności przekonań (eksploracja)[25]. Tę ostatnią zazwyczaj wiąże się z poszukiwaniem nowych doświadczeń (por. Luciw i in., 2013). Z perspektywy FEP nagrody zewnętrzna i wewnętrzna, czyli wybór pomiędzy eksploatacją a eksploracją, wiążą się wprost z normatywnym wymogiem minimalizacji rozbieżności, tj. dywergencji KL, pomiędzy rzeczywistymi danymi (prawdopodobieństwo *a posteriori*) a oczekiwanymi

[25] W Friston i in., 2015b, jest ona również określana jako wartość epistemiczna.

(prawdopodobieństwo *a priori*, tj. wewnętrzne parametry modelu generatywnego). Czym większa dywergencja KL, tym większa rozbieżność pomiędzy modelem czy stanami wewnętrznymi a środowiskiem, czyli stanami zewnętrznymi (por. Bogacz, 2017; Kiefer, Hohwy, 2018). Wartościowa (w sensie statystycznym) polityka działania maksymalizuje oczekiwaną użyteczność i wiarygodność modelu. Jest ona zależna od kontekstu i posiadanych przekonań. Z tego też powodu, w tym ujęciu, nie ma rozbieżności pomiędzy minimalizacją zaskoczenia a eksploracją świata. Dzieje się tak, ponieważ minimalizacja zaskoczenia, może prowadzić do maksymalizacji wiarygodności modelu, o ile podmiot wybierze takie zasady działania, jakie zwiększają prawdopodobieństwo np. poszukiwania nowych stanów rzeczy. Zatem to, w jaki sposób podmiot minimalizuje niepewność, jest zależne od tego, jakie sposoby działania wybierze w celu optymalizacji modelu. Mogą to być przykładowo takie zasady, jakie opierają się na aktywnej eksploracji i poszukiwaniu nowości[26].

Problematyczne dla tego ujęcia jest wyjaśnienie tego, w jaki sposób wybierane są zasady działań i co motywuje podmiot, że wybiera on np. takie zasady, jakie nakazują mu szukać nowości, a nie np. pozostać w domu? Do tego problemu wrócę niebawem. Teraz odniosę się do koncepcji uczenia się ze wzmocnieniem, która obok klasycznych teorii podejmowania decyzji jest drugim obiektem krytyki Fristona i jego współpracowników. Problematyka uczenia się jest kluczowa dla poruszanych tutaj zagadnień, bowiem pokazuje ona, w jaki sposób podmiot bądź inny system reaguje na otoczenie, zdobywa informacje i wykorzystuje możliwości przezeń oferowane. Problem ten, jak wykażę, jest istotny dla analizy mechanizmów normatywnych oraz ich ograniczeń.

5.3.2.3. Rodzaje uczenia się a aktywne wnioskowanie

Uczenie przez wzmocnienie (*reinforcement learning*), obok uczenia nadzorowanego (*supervised learning*) i nienadzorowanego (*unspervised learning*), jest jedną z podstawowych metod uczenia maszynowego, która oferuje wiele możliwych algorytmów opisujących procesy pozyskiwania oraz analizy danych przeznaczonych do ulepszania i rozwoju danego systemu[27]. Uczenie nadzorowane polega na tym, że system ma pewne określone wektory wejściowe oraz pożądane wartości wyjściowe. Nadzór polega na tym, że tworzy się funkcję, która odwzorowuje wejście systemu na jego wyjście. System na podstawie pewnych danych wyjściowych, określających

[26] Nowsze prace albo znacząco rozwijają propozycję Fristona i rozbudowują model wartości (m.in. Constant i in., 2019; Hesp i in., 2019), albo powołują się na psychologię ekologiczną i pojęcie afordancji i uzupełniają podstawową wykładnię FEP zaproponowaną przez Fristona (por. Bruineberg, Rietveld, 2014). W podrozdziałach *Predykcje skierowane na afordancje* i *Wnioskowanie aktywne jako wnioskowanie enaktywne* bliżej przyjrzę się tym propozycjom teoretycznym.

[27] Rodzajów uczenia maszynowego jest znacznie więcej, ale w niniejszych rozważaniach ograniczam się tylko do tych trzech.

jakieś możliwe pobudzenie, ma nauczyć się prawidłowej czy pożądanej reakcji bądź odpowiedzi. Zostaje wytworzony zbiór par zawierających pewien obiekt, który się uczy, oraz pożądaną przez nadzorcę odpowiedź. Wykorzystując taki zbiór, możliwe jest bezpośrednie policzenie błędu popełnionego przez algorytm uczący i wyliczenie konkretnych poprawek. Można powiedzieć, że dane wejściowe, do których system ma dostęp, są wykorzystywane do wyszukiwania zależności, służących do rozwiązania określonego problemu. Przykładem uczenia nadzorowanego jest rozwiązywanie testu przez uczniów na podstawie pewnego klucza. W odniesieniu do interesującej nas problematyki PP algorytmy uczenia nadzorowanego wydają się nieprzydatne, bowiem zgodnie z bazowym założeniem PP podmiot bądź system nieustannie musi działać w nieokreślonym, zaskakującym i niepewnym środowisku. Z tego też powodu część badaczy uważa, że odpowiednie algorytmy uczenia się może zaoferować metoda uczenia nienadzorowanego (por. Korbak, 2019a).

W uczeniu nienadzorowanym system nie ma żadnego „klucza interpretującego" dane. Musi samodzielnie wyszukać odpowiednie wzorce, zależności i relacje. Oznacza to, że nie ma on żadnych określonych danych na wejściu, które w jakimś sensie sugerują albo wskazują odpowiedź na wyjściu. Nabywa on zatem wiedzę nie po to, by zrealizować określone zadanie (wyznaczone np. przez projektanta albo nauczyciela, jak ma to miejsce w uczeniu nadzorowanym), ale uczy się w sposób spontaniczny. Podstawową motywacją systemu do uczenia się bez nadzoru jest to, że chociaż dane przekazywane algorytmom uczenia się są niezwykle bogate w strukturę wewnętrzną (np. obrazy, filmy i tekst), to cele i nagrody wykorzystywane w szkoleniu są nieliczne. Oznacza to, że to, czego uczy się system, polega nie tyle na zastosowaniu danych do określonych zadań, ile na samym ich zrozumieniu (Graves, Clancy, 2019). Uczenie bez nadzoru ma funkcję straty, która zależy jedynie od danych wejściowych modelu (i statystycznych wzorców, które wykazują) i nie jest zależna od jakichkolwiek informacji zewnętrznych. Badacze jako przykład nauki bez nadzoru wskazują takie zjawiska jak przewidywanie szeregów czasowych czy modelowanie językowe. W modelowaniu językowym uczenie nienadzorowane polega na tym, że wymaga się tutaj tylko liniowego uporządkowania słów w zestawie szkoleniowym (tj. pochodzących z sąsiadującego tekstu), bez konieczności przypisywania dodatkowych etykiet (Korbak, 2019a). Na przykład znając zdanie „W zawodach mogą brać udział zarówno dorośli, jak i …", możemy bez trudu przewidzieć, że w miejscu trzech kropek należy wstawić wyraz „dzieci". Takie dane (jak to przykładowe zdanie) mają większą moc predykcyjną, zatem stosunkowo prosto uczymy się rozwijania takich ciągów. Oczywiście można zasadnie zapytać, czy nie jest jednak tak, że umiejętność przewidzenia, że teraz powinno pojawić się słowo „dzieci", nie jest umiejętnością nabytą na drodze uczenia pod nadzorem (np. na lekcjach w szkole podstawowej)? Ta wątpliwość dowodzi tylko tego, że podział na uczenie nadzorowane i bez nadzoru nie jest ścisły i mocno

wiąże się z takimi kwestiami jak wyznaczone dla systemu zadanie do wykonania, dostęp do informacji[28].

Szczególnym przypadkiem uczenia maszynowego jest uczenie ze wzmacnianiem. Tutaj system nie ma jasno określonych danych wyjściowych, jednak z założenia otrzymuje on ocenę swoich działań w formie nagrody (wzmocnienie). Innymi słowy, można rozpoznać działania jako dobre lub złe, poprawne bądź niepoprawne. System ma się nauczyć działać w taki sposób, aby mógł maksymalizować wzmocnienie. Takie uczenie się jest o tyle trudne, że podmiot nie ma pełnej informacji o swoim środowisku, jak również wiedzy na temat skutków i konsekwencji swoich decyzji oraz działań. Wie jednak, że skuteczność jego działań jest zależna od maksymalizacji wzmocnień, które są jakąś formą korzyści dla niego (np. przyjemność, zaspokojenie głodu, utrzymanie bezpieczeństwa). W odniesieniu do tego rodzaju uczenia się trudno jest jasno określić, na czym polega kara za popełniony błąd. Karą może być oczywiście brak wzmocnienia, niemniej nie zawsze mamy do czynienia z takimi sytuacjami. Można wyobrazić sobie wieloetapowe działania, które mogą zakończyć się porażką (np. gra w szachy) i w których nie jest do końca jasne, kiedy popełniliśmy błąd. Szachy mogą zakończyć się (bo np. upłynął umówiony czas) bez jasnego werdyktu, kto wygrał, a kto przegrał. Z innej strony karą może być uszczerbek na zdrowiu i trudno o nim mówić po prostu jako o braku wzmocnienia.

W pewnym sensie uczenie ze wzmocnieniem jest metodą prób i błędów. System, aby mógł działać optymalnie w nieznanym środowisku, musi realizować zadania w określony sposób. Jednak sposób, w jaki system będzie działał, zależy od podjętych wcześniej decyzji. To znaczy, że jeśli wcześniej system popełniał błędy, teraz będąc w tym samym stanie, nie powinien podejmować takich akcji. Wiedząc, jak ocenione zostają jego decyzje, może starać się tak zmodyfikować swoje zasady działania, aby w kolejnych krokach otrzymać jak najmocniejsze wzmocnienie. Twierdzi się, że uczenie ze wzmocnieniem polega na znalezieniu optymalnej strategii decyzji, której modelem jest proces decyzyjny Markowa. W tym ujęciu wartością funkcji wzmocnienia jest zmienna losowa.

Nietrudno zauważyć, że pomiędzy uczeniem ze wzmocnieniem a koncepcją Fristona, opartą na maksymalizacji wiarygodności modelu na drodze minimalizowania niepewności, zachodzą istotne podobieństwa. Zwolennicy FEP twierdzą jednak, że ich koncepcja nie korzysta z rozwiązań uczenia maszynowego (por. Friston, Daunizeau, Kiebel, 2009; Friston, Adams, Montague, 2012; Friston i in., 2016; Schwartenbeck i in., 2013). Ich zdaniem nie potrzeba wprowadzać żadnych algorytmów uczenia maszynowego, w szczególności uczenia na podstawie wzmocnienia, aby wyjaśnić optymalność działań. Wystarczy jedynie odnieść się

[28] Uwaga ta również prowadzi do dalszej konkluzji, zgodnie z którą może być tak, że wszelkie procesy uczenia się, również te odniesione do PP, opierają się na łączeniu różnych metod. Czasem uczymy się pod nadzorem, czasem zaś działamy w takich warunkach, jakie wykluczają jakikolwiek nadzór. Nie oznacza to jednak, że nie wykorzystujemy wtedy wiedzy i umiejętności nabytych na drodze uczenia nadzorowanego.

do koncepcji, która wyjaśnia zachowania adaptacyjne i działania, wykorzystując postulat minimalizacji energii swobodnej. W tym ujęciu podmiot dostosowuje swoje stany wewnętrzne i działa w środowisku, aktywnie je eksplorując, ponieważ chce zminimalizować niepewność i zaskoczenie, których źródłem są nieznane stany tego środowiska. Poprzez aktywne działanie w tym środowisku podmiot uczy się zachowań adaptacyjnych i samokontroli. Jednocześnie wytwarza określone zasady działania, które są optymalizowane wraz ze zmianami stanów otoczenia. Działa zatem w taki sposób, aby jego przekonania (powiązane w hierarchiczną sieć bayesowską) z jednej strony były coraz bardziej koherentne, a z drugiej wiarygodne, w sensie bycia skutecznymi w przewidywaniu przyszłych stanów świata. Jest to możliwe poprzez wybór takich działań, jakie minimalizują niepewność, *resp.* oczekiwaną energię swobodną, maksymalizują wewnętrzne parametry modelu. W praktyce oznacza to, że nie trzeba zupełnie powoływać się na pochodzące z uczenia maszynowego i behawioryzmu pojęcia wzmocnienia, nagrody, kary czy wywodzące się z klasycznych teorii decyzji pojęcia użyteczności bądź wartości użytecznej. Podmiot bowiem uczy się działać na zasadzie kształtowania nawyków. Te zaś wyłaniają się w sposób naturalny, a nie nadzorowany i sztuczny z sekwencyjnej optymalizacji zasad działania podmiotu. Każde zachowanie ma określone aspekty odkrywcze (zachowania i działania epistemiczne) oraz eksploatacyjne (zachowania i działania pragmatyczne). Te pierwsze są wrażliwe na niepewność i ryzyko powiązane z nieokreślonością środowiska, te drugie zaś na poszukiwania nagrody. Jedne i drugie są odpowiedzialne za powstawanie nawyków (Friston i in., 2016).

5.3.3. Uwagi wobec ramy aktywnego wnioskowania

Wyjaśnienie optymalizacji działań i rozwiązania kompromisu eksploracja – eksploatacja w odniesieniu do ramy aktywnego wnioskowania i FEP nie jest wolne od słabości. Po pierwsze pojawia się poruszany tutaj problem z motywacją. Jeżeli przyjmie się propozycję Fristona i współpracowników, to należy stwierdzić, że ostatecznie podmiot sam w sobie wytwarza motywację. Mianowicie: do działania motywuje go to, co zwiększa wierzytelność jego przekonań i modelu świata. Co zatem zwiększa wierzytelność jego przekonań? Minimalizacja niepewności (por. Schwartenbeck i in., 2015). Jeżeli tak jest, to powraca tutaj problem ciemnego pokoju. Wiemy już jednak, że podmiot może przyjąć też takie zasady, jakie opierają się na aktywnej eksploracji i poszukiwaniu nowości. Nie będzie on zatem dążył do przewidywalnych i mało zaskakujących stanów, ale będzie w pewnym stopniu zwiększał stopień niepewności. Problematyczna jest jednak odpowiedź na pytanie dotyczące czynnika bądź czynników, które wpływają lub kształtują takie zasady działania. Mamy tutaj do czynienia z regresem w nieskończoność: potrzebę eks-

ploracji i poszukiwania nowości uzasadnia się poprzez odniesienie do takich zasad działania, jakie faworyzują taką potrzebę. Co zatem uzasadnia potrzebę wyboru takich zasad działania? Czy jest to jakieś uprzednie przekonanie podmiotu bądź ich zbiór? A w jaki sposób uzasadnić te przekonania? Czy przez odniesienie do jakichś wysokopoziomowych przekonań wyrażających potrzebę dążenia do nowości i zwiększania niepewności? Kwestie te są niejasne. Jednak na podstawie analizy najnowszej literatury (por. Veissière i in., 2020) można sądzić, że badacze aktywnego wnioskowania niebawem odpowiedzą na powyższe wątpliwości.

Kahneman i Tversky (1974, 1979) wykazali, że nie można satysfakcjonująco wyjaśnić procesów podejmowania decyzji i działania w warunkach niepewności, jeżeli przyjmie się postulaty teorii normatywnych. Co prawda ich krytyka była wymierzona w teorię oczekiwanej użyteczności, ale jest ona, jak twierdzę, zasadna w pewnym stopniu w odniesieniu do ramy aktywnego wnioskowania. Rama ta jako bayesowska zakłada subiektywną ocenę prawdopodobieństwa przez podmiot, podobnie jak ma to miejsce w teorii oczekiwanej użyteczności. Zdaniem Kahnemana i Tversky'ego wiele, jeżeli nie wszystkie, oceny i decyzje dotyczące działań opierają się na przekonaniach dotyczących prawdopodobieństwa zajścia pewnych zdarzeń czy pojawienia się pewnych stanów rzeczy. Zgodnie z ujęciami normatywnymi, działania w warunkach niepewności, czyli braku pewności w kwestii prawdopodobieństwa zajścia oczekiwanych sytuacji bądź konsekwencji wybranych działań itd., opierają się na pewnych ogólnych wymogach takich jak konieczność uwiarygodnienia naszych przekonań czy właśnie maksymalizacja oczekiwanej użyteczności. Autorzy ci wykazali, że z subiektywnym szacowaniem niepewności wiąże się wykorzystywanie przez ludzi heurystyk poznawczych. Kahneman i Tversky wymienili trzy podstawowe heurystyki: (1) heurystyka reprezentatywności – pojawia się wtedy, gdy ma się ocenić prawdopodobieństwo, że dany obiekt bądź stan rzeczy należy do określonej klasy bądź grupy; (2) heurystyka umysłowej dostępności – korzysta się z niej wtedy, kiedy trzeba ocenić częstość zachodzenia danego zdarzenia bądź wiarygodność zaistnienia określonego stanu rzeczy; (3) heurystyka dostosowania się korekty w stosunku do wartości wyjściowej – wykorzystuje się ją wtedy, gdy przeprowadza się prognozy numeryczne przy dostępności określonej wartości. Istnienie heurystyk nie dowodzi tego, że ludzie nie działają racjonalnie bądź optymalnie, ale uzasadnia tezę, zgodnie z którą to właśnie subiektywne oceny prawdopodobieństwa decydują o preferencjach podmiotów, a w następstwie o podejmowanych decyzjach, a nie, tak jak twierdzą zwolennicy ujęć normatywnych, jakieś ogólne preferencje czy zasady, które wyznaczają sposoby działania oraz stanowią normę je oceniającą i uzasadniającą. Heurystyki są skuteczne i spełniają wiele istotnych funkcji, niemniej ich istnienie i wykorzystywanie przez podmiot prowadzi do systematycznych oraz częstych i przewidywalnych błędów, nie tylko w ocenie prawdopodobieństwa, ale także w podejmowanych decyzjach.

W odniesieniu do ramy aktywnego wnioskowania oznacza to, że przy wyjaśnianiu możliwości działań należy wziąć pod uwagę kwestię heurystyk (por. podrozdział *Problemy z modelami bayesowskimi*). Są one, co potwierdzają liczne badania, w pewnych warunkach skuteczniejsze niż złożone obliczenia. Niektóre analizy z zakresu aktywnego wnioskowania odnoszą się do tego problemu (por. Constant i in., 2019; Veissière i in., 2020), niemniej istnieje pewne ogólne przekonanie, że sama struktura modelu generatywnego i dynamiczne zasady ją regulujące są eksplanacyjnie satysfakcjonujące[29].

Można bronić ujęcia Fristona, twierdząc np., że wiarygodność modelu nie tylko opiera się na jej wzmacnianiu przez działania, ale przede wszystkim na drodze samouwierzytelniania się (*self-evidencing*) (por. Hohwy, 2016, 2020), czyli odwołania się do wewnętrznej spójności i koherencji modelu. Niemniej w świetle badań Kahnemana i Tversky'ego nie jest to zadowalające kryterium, bowiem albo cały model jest koherentny, albo nie[30]. Niemożliwe jest posiadanie takich przekonań dotyczących oceny prawdopodobieństwa danych zdarzeń, jakie nie byłyby spójne z całą siecią przekonań żywionych przez daną osobę. Nie istnieje bowiem prosta procedura formalna, która pozwalałaby ocenić zgodność niektórych przekonań z całościowym systemem przekonań danego podmiotu (Kahneman, Tversky, 1979). Pozostaje zatem sprawdzenie ich w działaniu, co znów będzie opierało się na heurystykach.

Zwolennik *active inference* może odnieść się do, przywoływanego już wcześniej, pełnego twierdzenia klasowego. Mianowicie: można zamodelować sieć bayesowską w taki sposób, aby działała ona tak jak kombinacja heurystyk. W praktyce uczenia maszynowego taki zabieg nazywa się „regularyzacją". Polega on na dodaniu niektórych informacji lub ograniczeń w celu uniknięcia przeuczenia lub rozwiązania źle postawionego problemu. W odniesieniu do FEP oznaczałoby to, że minimalizacja wielu funkcji sprowadza się do minimalizacji jednej funkcji będącej sumą wielu funkcji. Rozwiązaniem jest zatem dodanie do funkcji kosztu kary za złożoność modelu. Zostaje zminimalizowana ważona suma dokładności i złożoności modelu. W ten sposób dwa ograniczenia składają się na jedną wartość. W języku FEP te dodatkowe ograniczenia bądź heurystyki możemy dopasować do wewnętrznych parametrów modelu, co oznacza, że minimalizacja energii swobodnej będzie minimalizowała również jakąś ich kombinację[31]. Zabieg ten jednak może mieć istotne

[29] Twierdzenie to należy traktować z pewnym dystansem, ponieważ rama aktywnego wnioskowania jest nieustannie modyfikowana i wzbogacana przez kolejne analizy. Twierdzi się (Ramstead), że obecnie (lipiec 2020) mamy do czynienia z trzecią falą aktywnego wnioskowania.

[30] Paul Thagard i Karsten Verbeurgt (1998) podkreślają, że istnieją określone metody aproksymacji koherencji modelu, jednak zwolennicy aktywnego wnioskowania nie powołują się na nie.

[31] Optymalizacja wielu celów funkcji $f(x)$ i $g(x)$ sprowadza się do optymalizacji jednego celu funkcji takiej, że $h(x) = a^*f(x) + b^*g(x)$. Należy tylko dobrać wagi a i b oraz wszystkie ograniczenia, jakie chce się ująć w ramach jednej funkcji kosztu, pozostając w obrębie wymogu minimalizacji FEP. W języku sieci bayesowskich mamy po prostu model hierarchiczny, z podsieciami reprezentującymi f i g.

ograniczenia, jeżeli otrzymana kombinacja heurystyk daje większą złożoność niż ta obliczana przez procesy decyzyjne Markowa. Chomsky (1957) pokazał, że procesy te nie pozwalają na modelowanie bardziej skomplikowanych zasad działania, np. języka naturalnego, którego gramatyki nie da się do nich sprowadzić. Język jest tak skonstruowany, że możemy generować wszystkie wypowiedzi, określając dozwolone sekwencje elementów najwyższego poziomu itd. Jeżeli zatem ludzkie procesy selekcji i podejmowania działań w określonych warunkach środowiskowych są równie złożone jak te opisywane przez Chomsky'ego, a z całą pewnością niektóre są, to zamodelowanie heurystyk i ich sprawna optymalizacja w obrębie sieci, może stanowić istotną trudność dla ramy aktywnego wnioskowania jako ramy wartościowej eksplanacyjnie.

Krytykę w podobnym duchu można przeprowadzić z perspektywy teorii uczenia maszynowego: (1) istnieje wiele badań, które wykazują, że algorytmy nauczania nadzorowanego można wykorzystać w hierarchicznych sieciach bayesowskich (por. Eckstein, Collins, 2019); (2) istnieją argumenty na rzecz tezy, zgodnie z którą dla wyjaśnienia działań w ramach FEP wystarczy przyjąć algorytmy uczenia nienadzorowanego (por. Korbak, 2019; Ofner, Stober, 2018); (3) wbrew zapewnieniom Fristona i jego zwolenników, jeżeli uzna się za wartość funkcji wzmocnienia maksymalizację precyzji przekonań, *resp*. wewnętrznych parametrów modelu, to wnioskowanie aktywne jako optymalizację działań można uznać za działanie oparte na uczeniu się ze wzmocnieniem, czyli za jeden z wariantów tego algorytmu; (4) wielu autorów zwraca uwagę na fakt, że aby dobrze wyjaśnić działania systemu w realnym środowisku, należy przyjąć szereg różnych algorytmów dotyczących tego, w jaki sposób system nabywa wiedzę i umiejętności. Powoli bowiem odchodzi się od przekonania, zgodnie z którym zadania wyznaczone przez inżyniera określonemu systemowi da się zrealizować tylko, wykorzystując jeden algorytm (Lukac, Kameyama, 2015). W odniesieniu do faktycznych działań w warunkach niepewności ta potrzeba wydaje się jeszcze bardziej widoczna. Należy zatem stwierdzić, że odpowiednią ramę wyjaśniającą działania można zbudować jedynie w odniesieniu do jakiejś hybrydowej koncepcji uczenia się.

5.4. Normatywne mechanizmy predykcyjne a ograniczenia

Badacze sztucznej inteligencji oraz uczenia maszynowego dobrze zdają sobie sprawę z wielu trudności, jakie narzuca problem motywacji. Po pierwsze samo zaprojektowanie funkcji nagradzania jest trudne i kosztowne, po drugie zaś nie stosuje się tego zabiegu, albo robi się to w bardzo ograniczony sposób, w odniesieniu do realnych procesów osadzonych w prawdziwym środowisku. Aby uczynić tego typu

próby wyjaśniania bardziej realistycznymi, badacze przyjmują dodatkowe ograniczenia, takie jak twierdzenie, że czasem ludzie i zwierzęta robią coś dla samej tej czynności, bez wyraźnej nagrody (np. kot bawiący się kłębkiem wełny albo osoba rysująca palcem po piasku). Samo jednak wyjaśnienie tych ograniczeń nastręcza wiele trudności. Wiemy, że jedną z najważniejszych motywacji jest poszukiwanie nowości czy chęć poznawania świata, która pozwala unikać nudy (por. Gottlieb, Oudever, 2018). Kluczowe dla wyjaśnienia tego typu ograniczeń czy zjawisk jest, zdaniem wielu badaczy, odwołanie się do tzw. wewnętrznej motywacji (*intrinsic motivation*). Wewnętrzne sygnały motywacyjne często prowadzą do szybszego uczenia się oraz wytwarzania stabilniejszych wzorców zachowania. W ten sposób zwiększają funkcję nagradzania (por. Guo i in., 2016). Problem jednak polega na tym, że wewnętrzne sygnały motywacyjne raczej uzupełniają, a nie zastępują nagrody ze środowiska zewnętrznego. Tym samym otwartym problemem pozostaje pytanie o to, skąd się biorą nagrody, które można rozumieć tutaj jako przyczyny sygnałów motywacyjnych (Juechems, Summerfield, 2019)? Trudność ta również pojawia się w badaniach nad aktywnym wnioskowaniem.

Celem analiz przeprowadzonych w tej części pracy jest wykazanie, że koncepcja mechanizmów normatywnych pozwala na wyjaśnienie wspomnianej trudności. Wpierw omówię formalną możliwość zastosowania proponowanego tutaj ujęcia do problemu motywacji i nagrody, który zbiorczo określam jako kompromis ciemnego pokoju, a następnie zaproponuję odpowiednie rozwiązanie.

5.4.1. Mechanizmy normatywne: motywacja a nastawienie

Przez mechanizmy normatywne rozumiem te, które spełniają określone i nietrywialne role przyczynowe w wyjaśnianiu danych zjawisk. Mówiąc o takich rolach w odniesieniu do normatywnych własności mechanizmów oraz funkcji, mam na myśli to, że są one przyczynami określonych działań danego organizmu bądź podmiotu w jego środowisku. Normatywność jest zatem predykatem, za pomocą którego wyjaśnia się pewne zjawiska w kategoriach mechanizmów i funkcji, które są ich przyczynami. Jeżeli są to mechanizmy i funkcje bayesowskie, jak ma to miejsce w PP, to w ich przypadku mówi się o normatywnym charakterze modelowania bayesowskiego, które opiera się na wyjaśnieniach zakładających istnienie takich funkcji i mechanizmów. Tym samym broni się realizmu w odniesieniu do tego rodzaju modelowania. W rozdziale *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego* wykazałem, że każdy mechanizm (normatywny, ale także nienormatywny) ma określone ograniczenia. Ograniczenia za Marrem rozumiem jako przyczynowe, skuteczne, czyli takie, jakie zapewniają niezbędne i wystarczające warunki dla funkcjonowania określonych procesów i mechani-

zmów (Marr, 1982, s. 111–116). W takim ujęciu ograniczenia w pewnym sensie są normami bądź zasadami, które określają granice i zasady realizacji takich, a nie innych procesów. Ograniczenia mogą być fizyczne, biologiczne, społeczne czy kulturowe. Ich analiza jest kluczowa dla wyjaśnienia konstytucji danego mechanizmu. Przykładowo: nie da się satysfakcjonująco wyjaśnić mechanizmu prowadzenia samochodu, jeżeli nie weźmie się fizycznych i symbolicznych ograniczeń związanych z ruchem drogowym (określone przepisy, znajomość znaków drogowych itd.). Jak dotąd nie przeprowadziłem pełnej analizy pojęcia ograniczenia. Zanim jednak jej dokonam, konieczne jest ustalenie pewnych ram badawczych.

Zgodnie z przekonaniem zwolenników ramy aktywnego wnioskowania, ale także wielu badaczy uczenia maszynowego, motywacja jest czynnikiem, który ma charakter wewnętrzny. Hasłowo rzecz ujmując: jeżeli chcemy wyjaśnić zachowania bądź działania danego podmiotu, należy odnieść się do zasad działania, które czynią jego postępowanie optymalnym. Podstawową motywacją w ujęciu Fristona i jego współpracowników jest potrzeba ciągłego zwiększania wiarygodności modelu świata, w który uzbrojony jest podmiot. Motywacja ma zatem charakter internalistyczny, co jest wzmocnione przekonaniem, że wartość jest dowodem dla modelu (por. Friston, Adams, Montague, 2012). Zwolennicy ramy aktywnego wnioskowania twierdzą, że należy odróżnić wartości zewnętrzne od wewnętrznych (epistemicznych). Minimalizacja oczekiwanej darmowej energii jest zatem równoznaczna z maksymalizacją wartości zewnętrznej lub oczekiwanej użyteczności (którą definiuje się w kategoriach wcześniejszych preferencji lub celów) przy jednoczesnym maksymalizowaniu przyrostu informacji lub wartości wewnętrznej. Tym sposobem, twierdzą, można rozwiązać dylemat eksploracyjno-eksploatacyjny. Wartość epistemiczna jest maksymalizowana do momentu, gdy nie nastąpi dalszy przyrost informacji (eksploracja). Eksploatacja zaś jest powiązana z maksymalizacją wartości zewnętrznej. Z ujęciem tym wiąże się problem definiowania motywacji w terminach internalistycznych. Zarówno wartość zewnętrzna, jak i epistemiczna jest zrelatywizowana do preferencji i przekonań podmiotu. Nie można tutaj odwołać się do wartości rozumianej jako swoiste *novum* dla podmiotu, czyli coś co wykracza poza jego oczekiwania i wyobrażenia, coś co jest dla niego zaskakujące i nieprzewidywalne, ale jednocześnie pociągające i przede wszystkim motywujące do dalszych działań. W odniesieniu do ujęcia mechanistycznego, wyjaśnienie działań jest tutaj formułowane na podstawie wskazania odpowiednich komponentów mechanizmu oraz ich organizacji. Jest to opracowywane w terminach bayesowskiego wnioskowania aktywnego. Badacze ci jednak nie powołują się na żadne ograniczenia zewnętrzne dla mechanizmu, zaś to, co można by uznać za ograniczenia wewnętrzne, jest ostatecznie sprowadzalne do wielu wewnętrznych parametrów mechanizmu, *resp.* jego organizacji. Nie twierdzę tutaj, że zwolennicy aktywnego wnioskowania wpisują się w mechanistyczne ujęcie wyjaśnień w kognitywistyce

i psychologii[32], ale można ich analizy odczytać przy użyciu pojęć zaczerpniętych z tego ujęcia. Z tej perspektywy zasadne jest stwierdzenie, że w ramach badań nad aktywnym wnioskowaniem ignoruje się istnienie zewnętrznych ograniczeń dla mechanizmów. Twierdzę, że motywację powinno się wyjaśniać w odniesieniu nie tylko do ograniczeń wewnętrznych, ale również zewnętrznych. Motywacja bowiem jest, na co wszyscy omawiani przeze mnie autorzy się zgadzają, istotnym elementem wyjaśniającym działania w świecie oraz prowadzącym do rozstrzygnięcia kompromisu pomiędzy eksplorowaniem a eksploatowaniem środowiska. Nie jest ona po prostu wyrazem potrzeb danego podmiotu, ile jest czynnikiem kontrolującym strategie działania ukierunkowane na poszukiwanie nowości. Wiąże się ona bowiem z przetwarzaniem informacji dotyczących poszczególnych przedmiotów w celu maksymalnego zoptymalizowania działania (Anselme, 2010, s. 292). Uważam, że to odniesienie do środowiska jako przyczyny ograniczeń zewnętrznych jest warunkiem koniecznym dla wyjaśnienia możliwości pojawienia się motywacji. *Coś*, co napotykam w swoim otoczeniu, może mnie zmotywować (w sensie wywołania odpowiedniej sekwencji działań ukierunkowanych na jakąś własność środowiska czy przedmiot w nim obecny) do podjęcia takich, a nie innych działań lub do ich zaniechania[33]. Świat bowiem nie tylko nas motywuje, ale także pokazuje ograniczenia dla celów wyznaczanych przez naszą motywację.

Problem, który analizuję, można wyjaśnić, wykorzystując znaną z metaetyki dyskusję nad tzw. argumentami konstytutywnymi (*constitutive arguments*). Najogólniej mówiąc, zwolennicy tych argumentów (por. Korsgaard, 1996, 2009; Rosati, 2003; Velleman, 2000) twierdzą, że umożliwiają one wyjaśnienie normatywnej mocy racjonalnych wymogów (takich jak omawiany tutaj wymóg maksymalizacji wierzytelności dowodu) poprzez wskazanie, że uznanie moralnych nakazów, zobowiązań, reguł itd. wynika z konstytutywnych cech podmiotu działania. Oznacza to, że nie jest konieczne przyjmowanie założenia o istnieniu zewnętrznych w stosunku do podmiotu własności czy przedmiotów normatywnych. Zgodnie z tym ujęciem normatywność niektórych wymagań można uzasadnić w sposób całkowicie naturalistyczny, bez odwoływania się do nieredukowalnych właściwości normatywnych znajdujących się poza podmiotem[34]. Zwolennicy stosowania argumentów konstytutywnych twierdzą, że rozpoznanie fundamentalnej roli, jaką odgrywa w działaniu podmiotowość, pozwala wskazać zarówno źródło ich motywacji, jak i wyjaśnić racjonalne podstawy uzasadnienia podjętych działań, które podmiot działający przypisuje sobie jako ich sprawca. Korsgaard twierdzi, że

[32] Faktem jest, że zwolennicy tego ujęcia piszą o organizacji mechanistycznej (por. Badcock, Friston, Ramstead, 2019), niemniej pojęcie mechanizmu jest tutaj używane w sposób niekrytyczny.

[33] Motywację do podjęcia działania lub do ich zaniechania Patric Anselme określa jako pozytywne i negatywne wzmocnienie (Anselme, 2010, s. 297). Wydaje się jednak, że jest to znaczne uproszczenie.

[34] Wśród konstytuwistów nie ma zgody co do tego, co jest własnością (bądź ich zbiorem) konstytutywną dla działań. Przykładowo David J. Velleman (2000) twierdzi, że własnością konstytutywną dla działań są określone cele, Korsgaard (2009) zaś utrzymuje, że jest to autonomiczna efektywność sfery wolitywnej.

wymogi określające działania mają podwójny charakter, bowiem są jednocześnie opisowe i normatywne. Są opisowe, ponieważ opisują określone działania, są zaś normatywne, ponieważ sankcjonują jako wadliwe przypadki działań, które nie są z nimi zgodne. Stwierdzenie zgodności bądź niezgodności działania wiąże się tutaj z autonomicznością podmiotu i możliwością autoidentyfikacji, czyli możliwością określenia siebie jako odpowiedzialnego sprawcy działania, co oznacza, że warunkiem bycia agentem jest identyfikacja z zasadą wyboru, na podstawie której się działa. W innym wypadku podmiot nie jest podmiotem, ale „szeregiem niepowiązanych ze sobą impulsów" (Korsgaard, 2009, s. 75–76).

Powyższa argumentacja jest, jak twierdzę, analogiczna w odniesieniu do strategii argumentacyjnej stosowanej w PP i w pewnym stopniu w ramach aktywnego wnioskowania. Tutaj wymóg długoterminowej minimalizacji błędu predykcyjnego, *resp.* maksymalizacji wierzytelności modelu, czerpie swoją moc normatywną z dynamiki modelu bayesowskiego uzasadniającego, które z zachowań lub działań staje się optymalne: coś jest ograniczone jedynie w odniesieniu do zasad i parametrów modelu, ale także do (bayesowskiej) reguły wnioskowania. Oznacza to, że normatywne wymagania aktywnego wnioskowania są wymaganiami dotyczącymi formowania przekonań (takich jak to, że jesteśmy wrażliwi na dowody lub pewne reguły wnioskowania). Ujęcie takie rodzi jednak pewne trudności. Pierwsza z nich dotyczy możliwości działań niezgodnych z normatywnym wymogiem maksymalizacji wierzytelności modelu: jeżeli bowiem wszelkie działania są formą maksymalizacji wierzytelności dowodu, to jak wyjaśnić możliwość błędnych, niepoprawnych albo po prostu niezgodnych z tą normą działań. Być może jest tak, jak utrzymują zwolennicy wnioskowania aktywnego, że wszelkie działania są *a priori* zgodne z tym wymogiem, ale jeżeli tak jest, to proponowana rama nie oferuje wyjaśnienia działań, ile ich redeskrypcje w terminach optymalizacji sieci Bayesa (por. podrozdział *Wnioskowanie aktywne jako wnioskowanie enaktywne*; Van Es, 2020).

Problem drugi wiąże się z podstawowym założeniem zwolenników argumentów konstytutywnych, według którego podmiot jest zobowiązany do przestrzegania określonych wymogów normatywnych. Można jednak wykazać pewną niespójność w przypadku podmiotu, który nie przestrzega normatywnych wymagań (por. Enoch, 2006). Zwolennicy tego stanowiska twierdzą, że gdy ktoś bierze udział w jakiejś grze, to jest zobowiązany do przestrzegania pewnych zasad, więc ignorowanie tych zasad jest sprzeczne z tym, do czego agent już się zobowiązał. Taki podmiot jest zatem irracjonalny, bowiem zarówno akceptuje ich moc normatywną, jak i jej zaprzecza. Jest to problematyczne, ponieważ podmiot musi przyjąć uzasadnienie dla akceptacji danego przekonania (np. że powinien opuścić ciemny pokój z powodu braku dostępu do światła i pożywienia), jak i móc uzasadnić jego odrzucenie. Zgodnie z tym ujęciem podmiot zachowujący się w ten sposób jest

niespójny albo irracjonalny. Nie odmówimy jednak racjonalności osobie, która z jakichś powodów łamie np. określany nakaz w imię tzw. „wyższych" czy „innych" wartości. David Enoch problematyczność tego ujęcia wiąże z przekonaniem, że racje czy przekonania, które uznaje podmiot za obowiązujące, nie mogą mieć swojego źródła poza tym podmiotem. Taki pogląd jest spójny z ramą aktywnego wnioskowania, ponieważ tutaj o optymalności danych działań i zachowań decyduje jedynie dynamika (normatywnego) modelu generatywnego, który jest samouwierzytelniający i oferuje imperatywy dla działań (Constant i in., 2019, s. 5). Z punktu widzenia przeciwników tego rodzaju argumentów podmiot jest niespójny i grozi mu irracjonalność (por. Williams, 2019). Oznacza to, że wyjaśnienie przestrzegania i nieprzestrzegania określonych wymogów oraz ich optymalizacja ma określone koszty poznawcze (Kahneman, Tversky, 1974, 1979) powiązane z naszymi ograniczeniami ewolucyjnymi i obliczeniowymi (problem NP-trudny). Twierdzę, że przyjęcie ramy PP i w pewnym stopniu aktywnego wnioskowania[35] wymaga uznania, że podmioty muszą działać w sposób spójny pomimo tego, że czasem bywają niespójne, rozdarte pomiędzy różnymi alternatywami i pragnieniami, czy wewnętrznie skonfliktowane itd.[36] Trzeba zatem przyjąć pewną wersję założenia o idealnym lub racjonalnym podmiocie poznającym.

Jeszcze innym problemem jest możliwość wyjaśnienia, jedynie z punktu widzenia organizacji systemu, normatywnej różnicy pomiędzy udanym działaniem prowadzącym do sukcesu a przypadkowym zachowaniem albo jakimś błędem lub pomyłką zrobioną przy okazji danego działania (por. Bickhard, 2016). Przykładowo, czy stłuczenie przeze mnie filiżanki podczas nalewania herbaty jest formą optymalizacji mojego systemu przekonań? Dlaczego nalewanie herbaty ma maksymalizować wierzytelność mojego modelu, a stłuczenie filiżanki już nie? Jak odróżnić tutaj bycie podmiotem działania od bycia przyczyną pewnego działania?

Na potrzeby niniejszych badań wprowadzę istotne rozróżnienie pomiędzy motywacją zewnętrzną, która jest motywacją w sensie ścisłym, a motywacją wewnętrzną, którą od tej pory będę określał mianem „nastawienia"[37]. Twierdzę, że dopiero para „motywacja – nastawienie" może umożliwić nam wyjaśnienie działań podmiotu w środowisku. Motywacja jest ściśle powiązana z odbiorczością i skierowana na to, co motywuje podmiot z zewnątrz. Nastawienie zaś wiąże się z ogólnym kierunkiem działania wyznaczanym przez preferencje, przekonania i możliwości określonego podmiotu. Nastawienie jest zatem tym, co zwolennicy

[35] Piszę „w pewnym stopniu", ponieważ pojawiają się obecnie liczne próby rozwiązania tych trudności na gruncie tej ramy badawczej.

[36] Dobrym przykładem jest zjawisko dysonansu poznawczego, które polega na poczuciu niespójności pomiędzy własnymi przekonaniami lub pomiędzy przekonaniami i podejmowanymi działaniami (por. Festinger, 1957; Kaaronen, 2018).

[37] Na temat znaczenia pojęcia nastawienia w zagadnieniach epistemologicznych por. Krokos, 2006.

aktywnego wnioskowania określają mianem „motywacji wewnętrznej”[38]. W tym ujęciu określone zasady działania są regulowane zarówno przez nastawienie, jak i motywację podmiotu. Zatem pełne ich wyjaśnienie zakłada konieczność zrozumienia sposobu, w jaki środowisko, tj. ograniczenia zewnętrzne, konstytuuje motywację. Pojęć tych jednak nie należy rozumieć w sensie psychologicznym, ale informacyjno-statystycznym, jako określone tendencje do zachowywania się i działania wyznaczane przez parametry modelu generatywnego, określone stany organizmu oraz ich relacje ze stanami środowiska. Chodzi o to, że większość zmian w środowisku wywołana jest zmiennymi, których rozkładu prawdopodobieństwa nie znamy. W tym też sensie niepewność nie jest własnością stanów rzeczy, ale pewnym stanem podmiotu, który jest skonfrontowany ze środowiskiem. Niepewność wiąże się z przekonaniem, że zwiększa się prawdopodobieństwo pojawienia się potencjalnych błędów predykcyjnych, zatem zmniejszenie tego prawdopodobieństwa jest kluczowe dla przetrwania organizmu i skutecznego działania w świecie. Podmiot zatem powinien minimalizować bądź optymalizować dywergencję K-L, czyli różnicę pomiędzy prawdziwym rozkładem prawdopodobieństwa *a posteriori* w odniesieniu do informacji zmysłowej a rozkładem przybliżonym obliczanym przez model generatywny. W jaki sposób dywergencja K-L może zostać zoptymalizowana? Przez zbliżenie rozkładu przybliżonego do prawdziwego, czyli przez zwiększenie wierzytelności wewnętrznych parametrów modelu, tj. przez nabycie nowej wiedzy. Z tego też powodu rolą motywacji jest ukierunkowanie podmiotu na nowość, czyli zwiększenie potrzeby eksploracji środowiska. Motywacja jest zatem tym czynnikiem, który sprawia, że podmiot nie chce biernie pozostać w ciemnym pokoju. Trzeba zatem stwierdzić, że niepewność nie tylko ma potencjalnie szkodliwe działanie dla podmiotu, ale jest także, a być może głównie, własnością motywacyjną (Anselme, 2010). Z nastawieniem mamy analogiczną sytuację, z tą różnicą, że nastawienie jest ukierunkowane i niejako motywowane od wewnątrz, co oznacza, że dotyczy ono optymalizacji rozkładów prawdopodobieństwa powiązanych z przekonaniami i preferencjami podmiotu[39].

W odniesieniu do przyjętego w tej pracy sposobu mówienia motywacja i nastawienie określają własności mechanizmów normatywnych. Mam na myśli to, że realizacja określonej funkcji predykcyjnej zależna jest od tak, a nie inaczej ukształtowanych (w sposób zewnętrzny i wewnętrzny) parametrów modelu generatywnego uzbrojonego w dany mechanizm. Oznacza to, że pełne wyjaśnienie natury mechanizmów normatywnych zakłada również wyjaśnienie ich ograniczeń

[38] Zatem dopiero para nastawienia i motywacji określa w sposób pełny to, co jest motywacją. Nie da się zatem jej wyjaśnić, odnosząc się tylko do nastawienia, tj. motywacji wewnętrznej.

[39] Friston i in., 2012b, zwracają uwagę na rolę dopaminy w procesach wyboru optymalnych działań. W proponowanym ujęciu zmiany dopaminergiczne odpowiedzialne są za przypisywanie precyzji błędom predykcyjnym. Ma to mieć duże znaczenie w zrozumieniu procesów motywacyjnych. Por. także FitzGerald, Dolan, Friston, 2015.

zewnętrznych (a czasem również wewnętrznych), które to ograniczenia kształtują (przynajmniej niektóre) z parametrów wewnętrznego dla danego podmiotu modelu świata oraz wpływają na rozkład prawdopodobieństwa przestrzeni, w której realizowany jest dany mechanizm (Craver, 2007, s. 247). Twierdzę, że takie wyjaśnienie pozwoli również odpowiedzieć na pytania dotyczące możliwości działań w świecie oraz rozwiązać kompromis dotyczący wyboru pomiędzy działaniami eksploracyjnymi a eksploatacyjnymi.

Pełne wyjaśnienie motywacji zakłada zrozumienie tego, czym są ograniczenia dla mechanizmów. Nasze dotychczasowe analizy tego pojęcia nie są zadowalające. Jego wyjaśnieniu poświęcę najbliższe rozważania.

5.4.2. Ograniczenia

Pojęcie ograniczenia pochodzi z biologii teoretycznej, a zaproponował je Howard Pattee (1968, 1972). Jego zdaniem wskazanie ograniczeń dla funkcjonowania danego systemu ma zapewnić lepsze zrozumienie jego funkcjonowania. Pattee odróżnia ograniczenia od praw. Te ostatnie są konieczne i nie można ich uniknąć bądź ominąć. Ograniczenia przeciwnie: często są przypadkowe i względne. W dynamicznych opisach układów żywych odróżnienie to jednak traci na ostrości. Ograniczenia, w odróżnieniu od praw natury, muszą być konsekwencją określonych struktur materialnych, takich jak cząsteczki, membrany czy np. maszyny. Struktury te są statyczne, czyli w pewnym stopniu zależne od praw natury, jednak ich zachowania można tylko wyjaśnić poprzez wskazanie ich ograniczeń, które są zależne od czasu. Z tego to powodu Pattee określa je mianem „reguł" (por. także Marr, 1982, s. 22–23; podrozdział *Wyjaśnienie mechanizmów normatywnych a ograniczenia*). System i jego złożoność opisuje się w odniesieniu do wielu poziomów ograniczeń w zależności od tego, jak dokładnie chcemy dany system wyjaśnić. Można również, co podkreśla Pattee, mówić nie tyle o regule, ile o regulacji, nie zmienia to jednak istoty sprawy. Chodzi bowiem o to, że system ma możliwość zachowywać się w taki, a nie inny sposób. Zaś owo zachowanie można skorelować ze sterującą zmienną wejściową lub sygnałem o określonej dynamice wyjściowej, zgodnie z daną regułą. Dlaczego wskazanie ograniczeń na różnych poziomach organizacji systemu jest takie istotne? Dlatego że w przypadku procesów pomiarowych, rejestracyjnych, klasyfikacyjnych, percepcyjnych, decyzyjnych czy informacyjnych pewna liczba możliwych zachowań na jednym poziomie opisu jest zmniejszana czy właśnie ograniczana przez określoną regułę (ograniczenie) z wyższego poziomu opisu. Poza tym nie da się mówić o jakimś zdarzeniu bądź zjawisku jako możliwym i niemożliwym na tym samym poziomie opisu (Pattee, 1972, s. 250). Przykładowo: na jakimś niższym, prostszym poziomie określone zachowania są

możliwe, ponieważ ograniczają one możliwość zachowania bądź działania, np. podjęcia decyzji na poziomie wyższym. Zaś na tym wyższym poziomie określone reguły z poziomu niższego są uwierzytelniane, tzn. okazuje się, które z nich są faktycznie realizowane, a które nie[40].

Ogólnie rzecz ujmując, ograniczenia są tym, co zmniejsza stopień swobody danego systemu w odniesieniu do zmienności lub możliwości zmiany jego parametrów, komponentów i zachowań (Umerez, Mossio, 2016). Ograniczenia grają istotną rolę zarówno w wyjaśnieniach mechanistycznych, jak i podejściu dynamicznym. Jason Winning i William Bechtel (2018) słusznie zwracają uwagę na to, że mechanicyści stosunkowo późno zauważyli konieczność uwzględnienia roli ograniczeń w wyjaśnianiu zjawisk. Ich zdaniem można scharakteryzować mechanizmy w kategoriach ich ograniczeń. Strategia taka przynosi pożądane efekty np. w odniesieniu do wyjaśnień niektórych mechanizmów neuronalnych (por. Bechtel, 2019). Zwolennicy ujęcia dynamicznego (por. Rączaszek-Leonardi, 2012) sugerują, że badania i opis ograniczeń można postrzegać jako substytut perspektywy mechanistycznej[41]. Byłoby to korzystne m.in. z powodu uniknięcia części obciążeń teoretycznych związanych historycznie z językiem opisującym pracę maszyn i mechanizmów. Należy jednak podkreślić, że opis i wyjaśnienie zjawisk w wielu naukach cały czas opiera się na wskazywaniu i rekonstrukcji ich mechanizmów (por. Machamer, Darden, Craver, 2011).

W Craverowskiej (2007) wersji teorii wyjaśnień mechanistycznych przestrzeń możliwych mechanizmów, których opis stanowi wyjaśnienie określonych zjawisk, definiowana jest przez takie, a nie inne przedmioty (*entities*), własności, aktywności oraz ograniczenia, które określają organizację mechanizmu (np. grawitacja, ilość dostarczanej energii, emisja światła). Przez ograniczenie Craver rozumie takie odkrycie, jakie albo kształtuje granice przestrzeni prawdopodobnych mechanizmów, albo zmienia rozkład prawdopodobieństwa w tej przestrzeni, tzn. zmienia prawdopodobieństwo, że jakiś punkt lub region przestrzeni dokładnie opisuje faktyczny mechanizm. Niektóre ograniczenia wykluczają przykładowo określone regiony tej przestrzeni, pokazując, że pewien zestaw możliwych mechanizmów jest niemożliwy ze względu na ich organizację oraz komponenty. W tym sensie ograniczenia wyznaczają określone ramy do opisu mechanizmów. Opis mechanizmu można zatem postrzegać jako proces kształtowania i ograniczania przestrzeni prawdopodobnych mechanizmów (Craver, 2007, s. 247–248). Miłkowski zwraca uwagę na to, że istnieją też takie ograniczenia, których nie można uznać za odkrycia, a raczej za normy bądź zasady metodologiczne (Miłkowski, 2016a, s. 50). Jako

[40] Pattee zwraca uwagę na fakt, że takie ujęcie ograniczeń jest warunkowane perspektywą badawczą. Należy jednak odróżnić strategię eksplanacyjną, która faworyzuje pewne ograniczenia kosztem innych, od faktu, że te ograniczenia istnieją i określają dany organizm czy strukturę.

[41] Wbrew klasycznym podziałom Carlos Zednik (2008) przekonująco pokazuje, że wyjaśnienie dynamiczne można ostatecznie uznać za rodzaj wyjaśnień mechanistycznych.

przykład tego typu normatywnych ograniczeń można odnieść się do niektórych zasad metodologicznych wskazanych przez Wesleya Salmona (1984). Przykładowo: zwykłe sekwencje czasowe nie są wyjaśniające. W praktyce eksplanacyjnej oznacza to, że model przyczynowy mechanizmu powinien być ograniczony przez tego typu zasadę, aby można było o nim mówić jako w pełni wyjaśniającym (Craver, 2007, s. 26–27). W tym sensie modele mechanizmów ograniczają nie tylko ustalenia (tj. ograniczenia) empiryczne, ale także niektóre zasady lub normy metodologiczne (np. konwencje reprezentacyjne stosowane do tworzenia symulacji komputerowych) itd. (Miłkowski, 2016a, s. 50)[42].

Zwolennicy psychologii ekologicznej podkreślają jeszcze inne znaczenia pojęcia ograniczenia, które można w ich ujęciu rozumieć również jako wymuszenie (por. Dotow, Nie, de Wit, 2012). Przykładowo Anthony Chemero (2009), chcąc wyjaśnić, czym są ograniczenia i jak należy rozumieć informację, powołuje się na semantykę sytuacyjną Jona Barwisa i Johna Perry'ego (1981, 1983). Koncepcja ta wiąże się ściśle z ideą głoszącą konieczność powiązania rozważań ontologicznych z semantyką. Semantyka sytuacyjna dąży do zdefiniowania funkcji semantycznych, które pozwolą wyjaśnić sposób, w jaki znaki mogą stanowić nośnik informacji o świecie zewnętrznym. Chodzi mianowicie o to, w jaki sposób jedna sytuacja może wskazywać na inną. Z tego właśnie powodu informacja jest tutaj definiowana w terminach przestrzeni stanów dla danego systemu. Punktem wyjścia tak rozumianej semantyki jest formalne pojęcie sytuacji: model przepływu informacji ma abstrahować od uposażenia i zróżnicowania podmiotów, a także od różnych możliwych form reprezentacji. W przypadku teorii ekologicznej owa „abstrakcja od form reprezentacji" ma zasadnicze znaczenie metateoretyczne. „Coś" jest dla podmiotu sytuacją, o ile jest on przyzwyczajony do danego zdarzenia lub oswojony z nim. Oznacza to, że jest on w stanie rozpoznać pewne stałe elementy w zmiennym środowisku. Sytuacja jest zatem zawsze sytuacją dla kogoś. Chemero rozważa następujący przykład: wyobraźmy sobie zamkniętą puszkę piwa stojącą na stole. Jest ona pewną sytuacją *A*. Sytuacja *A* zawierająca informację *a*, prezentuje sytuację *B*, która informuje (*b*), że w puszce znajduje się piwo (Chemero, 2009, s. 117). Obecność jednej sytuacji bezpośrednio wskazuje na istnienie innej, która zawiera jakąś inną konkretną informację. Informacja *b* jest, co podkreśla Chemero, bezpośrednio zawarta w sytuacji *A*. Nie oznacza to jednak, że sytuacja *B* zawsze zachodzi, kiedy zachodzi sytuacja *A* (Chemero, 2009, s. 119)[43]. Istnieje fizyczna możliwość, że puszka jest pusta pomimo tego, że została zamknięta[44].

[42] Przykładowo takimi ograniczeniami dla modeli decyzyjnych będą heurystyki badane przez Kahnemana i Tversky'ego.

[43] Chemero w tym miejscu powołuje się na analizy Millikan.

[44] Analiza zasadności takiego rozumowania wykracza poza cele wyznaczone w tej pracy. Istnieje jednak duża dyskusja w literaturze nad tym problemem.

Kluczowa dla semantyki sytuacyjnej jest kategoria ograniczania (*constraint*), która rozjaśnia powyższy przykład puszki piwa. Ograniczenia w semantyce sytuacyjnej oznaczają możliwe powiązania pomiędzy sytuacjami. Ograniczeniami są m.in. prawa przyrody, konwencje i zasady logiczne, reguły językowe, zależności empiryczne oraz wszelkie inne regularności. Rozważmy następujący przykład: zdanie „Dym oznacza ogień" wyraża pewne ograniczenie rozpoznawane przez agentów takich jak człowiek. Co to znaczy? Jeżeli *S* jest typem sytuacji, w której obecny jest dym, a *S'* jest typem sytuacji, w której obecny jest ogień, to podmiot może uzyskać informację na podstawie obserwacji dymu, że gdzieś się pali.

Na poziomie językowym ograniczenie pojawia się w następującym wyrażeniu: „OGIEŃ *oznacza* ogień", co znaczy, że pojęcie OGIEŃ *oznacza* zjawisko fizyczne, jakim jest ogień (Devlin, 1995, s. 12–14). Różne sytuacje mogą wskazywać na jedno i to samo ograniczenie, które ma charakter informacyjny. Rozpoznanie tego typu ograniczeń pomaga podmiotom podejmować odpowiednie decyzje na podstawie określonych wnioskowań. Ograniczenia wiążą się zatem ściśle z poznawczym i behawioralnym dostosowaniem się danego organizmu do konkretnej sytuacji, osadzonej w niepewnych i zmiennych warunkach. Organizm niejako ogranicza dopływ odpowiednich informacji w celu określenia konkretnej sytuacji, czyli rozpoznania pewnej użytecznej, z jego punktu widzenia, konfiguracji przedmiotów (Chemero, 2009, s. 117, 216). Hasłowo rzecz ujmując: użyć informacji, to wiedzieć, z jakimi ograniczeniami ma się do czynienia, czyli działać ze względu na nie.

Ekologiczne podejście do ograniczeń, jeżeli nie wiązać go ze szczególnymi rozwiązaniami ontologicznymi i epistemologicznymi dotyczącymi zagadnień reprezentacji czy obliczeń, wydaje się nie stać w sprzeczności z podejściem mechanistycznym. Mamy tutaj raczej do czynienia z ujęciem ograniczeń z perspektyw wysokopoziomowych mechanizmów, percepcyjnych i kategoryzacji i nie tyle stanowi ono kontrpropozycję wobec mechanicyzmu, ile jego uzupełnienie. Intuicja ta koresponduje zresztą z pojęciem ograniczenia zaproponowanym przez Marra, na które powoływaliśmy się w rozdziale *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego*. Przypomnijmy: zdaniem tego autora ograniczenia są normami bądź regułami, które określają granice oraz sposób, w jaki są realizowane określone procesy. Dzięki nim badany proces bądź mechanizm jest realizowany w określonym środowisku przez odpowiedni system je realizujący. Są to także pewne fakty dotyczące świata realnego (Shagrir, 2010, s. 489), jak i również społeczno-kulturowego, ponieważ mówiąc o ograniczeniach nałożonych np. na maszynę drukarską, mamy nie tylko do czynienia z określonymi prawami (np. optyki), ale pewnymi faktami dotyczącymi znaczenia znaków, które pojawiają się na druku itd. Na społeczny i materialny wymiar ograniczeń zwraca uwagę Donald Norman.

5.4.2.1. Znaczenie ograniczeń dla działania

Norman (2018, s. 145–153) odróżnia ograniczenia fizyczne, kulturowe, semantyczne i logiczne. Te pierwsze wiążą się z określonymi własnościami świata materialnego i zmniejszają liczbę potencjalnych działań. Na przykład samochód ciężarowy może wjechać do tunelu o określonych wymiarach przestrzennych. Ograniczenia kulturowe odnoszą się do zbioru działań czy zachowań, które są dopuszczalne w określonych sytuacjach. Ograniczają one w inny sposób potencjalne działania niż robią to ograniczenia fizyczne, ale ich nieprzestrzeganie niesie ze sobą określone konsekwencje. Przykładowo: pewną normą kulturową jest ruch prawostronny na drogach w Polsce. Niezastosowanie się do takiego rodzaju ograniczenia może skutkować spowodowaniem wypadku. Aby zatem wyjaśnić mechanizm skutecznego poruszania się po mieście, należy odnieść się do reguły ruchu prawostronnego. Ograniczenia semantyczne pojawiają się w takich sytuacjach, w jakich znaczenie (nie tylko językowe) zmniejsza pulę możliwych działań bądź zachowań. Ten rodzaj ograniczeń odwołuje się do wiedzy. Prawidłowe poruszanie się samochodem po mieście zakłada znajomość takich ograniczeń semantycznych jak znaki drogowe. Ograniczenia logiczne zasadzają się zaś na logicznych zależnościach pomiędzy różnymi faktami (sytuacjami, przedmiotami czy ich własnościami). To, jak zostanie zaprojektowany garaż, zależy od wielu czynników. Na przykład jeżeli droga do domu biegnie po jego lewej stronie, to garaż powinno się postawić również po lewej stronie domu. Można oczywiście postąpić inaczej, ale z punktu widzenia lokalizacji budynku mieszkalnego nie jest to działanie zasadne (może być zaś zasadne z powodu istnienia jakiegoś innego logicznego ograniczenia)[45].

Na podstawie powyższych analiz należy stwierdzić, że ograniczenia są istotnym komponentem mechanizmów, które mogą służyć za wyjaśnienia odpowiednich sytuacji, działań, zjawisk bądź procesów. Przykład poruszania się samochodem w przestrzeni miejskiej dobitnie to pokazuje. Jeżeli nie weźmie się pod uwagę wielu możliwych i faktycznie wykorzystywanych ograniczeń podczas jazdy samochodem, to wyjaśnienie tego zjawiska albo jest trywialne, albo schematyczne, czyli ostatecznie nie jest to dobre wyjaśnienie.

Teraz mogę wyjaśnić, jaki wpływ na motywację oraz nastawienie mają ograniczenia. To, co podmiot napotyka w ustrukturyzowanym przez ograniczenia środowisku, w sposób konstytutywny, a nie tylko przyczynowy, wpływa na jego motywację, której celem jest zmniejszenie niepewności podmiotu. Pewne fizyczne, społeczne, symboliczne czy kulturowe własności świata ograniczają pulę możliwych działań podmiotu, wykluczając jedne opcje, a wskazując inne. Przykładowo: jeżeli znajdę się w lesie w pobliżu dzika, to własności tego ekosystemu zmniejszą możliwą ilość działań, które mogę podjąć. Mogę zatem wspiąć się na drzewo,

[45] Norman oprócz ograniczeń omawia jeszcze m.in. afordancje.

o ile jest ono odpowiednio duże i wytrzymałe oraz wtedy, gdy pozwala mi na to moja waga i kondycja. Mogę uciekać przed siebie i spróbować skryć się w leśnej gęstwinie. Wszystkie te możliwości wyznaczają odpowiednie ograniczenia. Jedne z nich mnie motywują, tj. zmniejszają niepewność związaną z przewidywaniem konsekwencji zaatakowania przez dzika (np. wizja skrycia się na wysokim drzewie, sprowadzalna do takich, a nie innych ograniczeń tego przedmiotu), inne zaś demotywują (wizja upadku z drzewa z powodu moich ograniczeń kondycyjnych) itd. Analogicznie sprawa wygląda z nastawieniem (tj. motywacją wewnętrzną): to, w jaki sposób jestem gotowy zareagować na niebezpieczeństwo, zależy od tego, jakie przyjmę nastawienie. Mogę zatem starać się np. przestraszyć dzika albo od razu uciec. Nastawienie w tym przypadku jest ograniczane przez moją znajomość lasu, wiedzę o własnych możliwościach fizycznych, przekonaniach i wiedzę na temat dzików itd. Skuteczne działanie jest więc w tym ujęciu zależne od wiedzy dotyczącej ograniczeń obecnych w środowisku, które współkształtują nastawienie oraz motywację. Mogę oczywiście, jak twierdzą zwolennicy aktywnego wnioskowania, odwołać się do motywacji i nastawienia budowanych w odniesieniu do wymogu zwiększenia wiarygodności mojego modelu świata, niemniej może się okazać, że spotkanie z dzikiem będzie dla mnie na tyle niespodziewane i nieprzewidywalne, że w radykalnym przypadku, zamiast oczekiwanej maksymalizacji moich przekonań, stracę własne życie. Uwaga ta stanie się jaśniejsza, gdy na nowo przyjrzę się, poruszanemu przez Fristona i jego zwolenników, problemowi wartości. Zanim to jednak uczynię, konieczne jest przeniesienie tych analiz na grunt PP.

5.4.2.2. Ograniczenia w przetwarzaniu predykcyjnym

Na początek należy zwrócić uwagę na pewną dwuznaczność, która może być uznana za istotną słabość omawianej tutaj koncepcji ograniczeń. Chodzi mianowicie o pewien pozorny błąd kategorialny. Mianowicie mówi się tutaj o eksplanacyjnym znaczeniu ograniczeń w wyjaśnieniu określonych mechanizmów bądź zjawisk; z innej strony zaś stwierdza się, że ograniczenia są tym, co w sposób konstytutywny kształtuje te mechanizmy lub zjawiska. Czy nie mamy tutaj do czynienia z pomieszaniem tego, co jest wyjaśniane, z tym, co wyjaśnia? Twierdzę, że nie, ponieważ znaczenie i ważność eksplanacyjna ograniczeń w wyjaśnieniach jest uwarunkowana faktyczną rolą przyczynową, jaką ograniczenia grają w takich, a nie innych zjawiskach, które chcemy wyjaśnić. Innymi słowy, wyjaśniamy mechanizm, odwołując się m.in. do komponentu ograniczeń, ponieważ same mechanizmy zawierają taki komponent (por. Craver, 2007[46]). Po tych wstępnych uwagach, mogę przejść do omówienia roli ograniczeń w PP.

[46] Craver (2007, s. 249) wskazuje m.in. następujące rodzaje ograniczeń dla mechanizmów: międzypoziomowe, przestrzenne (np. rozmiar, kształt), czasowe (np. porządek, trwanie).

Nietrudno zauważyć, że koncepcję wielopoziomowych ograniczeń zaproponowaną przez Pattee można wiązać z hierarchiczną strukturą modelu generatywnego. Nie jest kontrowersyjne stwierdzenie, że informacje o potencjalnym błędzie predykcyjnym pochodzące z niższych warstw modelu stanowią ograniczenie dla generowanych na wyższych poziomach modelu predykcji, które jednocześnie minimalizują, czyli ograniczają wpływ potencjalnych błędów predykcyjnych na parametry samego modelu, jak i działań podejmowanych pod wpływem wygenerowanych predykcji. W tym przypadku jednak mówienie o ograniczeniach niewiele wnosi do już wypracowanego modelu PP. Inaczej sprawa wygląda, kiedy chce się wyjaśnić wpływ informacji pochodzącej z otoczenia na parametry modelu i generowane przezeń predykcje (oraz w efekcie na podejmowane działania). Przytoczmy już omawiamy wcześniej przykład wykrywania krawędzi. Naturalnym fizycznym ograniczeniem jest lokalizacja przestrzenna danego obiektu, który jest przyczyną zmiany intensywności światła (Marr, 1982, s. 68–90). Innym przykładem jest proces rozpoznawania wzorców, dla którego naturalnym ograniczeniem mogą być regularne lub nielosowe zmienne w porządku określonych danych (por. Dennett, 2008; Haugeland, 1993, s. 58; podrozdział *Rzeczywiste wzorce w przetwarzaniu predykcyjnym*). Na wyższych poziomach organizacji modelu generatywnego również mamy do czynienia z ograniczeniami, które wpływają na taki, a nie inny kształt sygnału zmysłowego. Z tej perspektywy należy stwierdzić, iż pomimo faktu, że zgodnie z PP nie mamy bezpośredniego dostępu do źródeł i przyczyn sygnału, to świat nie jest w sposób jednorodny niedookreślony. Przeciwnie, jest on już wstępnie ustrukturyzowany właśnie przez to, że stanowi naturalne ograniczenie dla informacji, którą system poznawczy odbiera i przetwarza. W środowisku kulturowo-społecznym sprawne działanie opiera się na znajomości i umiejętności poruszania się wśród wielu ograniczeń, nie tylko fizycznych, ale przede wszystkim kulturowych, semantycznych i logicznych.

Pełna analiza roli, jaką grają ograniczenia w PP znacznie wykracza poza ramy przyjęte dla tych rozważań, niemniej muszę się do niej odnieść, bowiem to dopiero istnienie ograniczeń oraz rozpoznanie ich faktycznego znaczenia dla konstytucji motywacji oraz przekonań podmiotu pozwoli wyjaśnić kompromis eksploracyjno-eksploatacyjny i rozwiązać problem ciemnego pokoju.

W odniesieniu do problemu mechanizmów i funkcji normatywnych zagadnienie ograniczeń jest jednym z najważniejszych. W rozdziale czwartym (por. podrozdział *Przetwarzanie predykcyjne a mechanistyczny model wyjaśniania naukowego*) zostało wstępnie wykazane, że, po pierwsze, jednym z warunków, który musi spełnić funkcja normatywna jest przyczynianie się do utrzymania stabilności organizmu (w taki sposób, że kształtuje ona i ogranicza możliwe relacje, procesy i zachowania danego systemu); po drugie zaś, że normatywność predykcji determinowana jest przez określone role funkcjonalne w modelu generatywnym (m.in.

wybór i kierowanie działaniami), ale także, co jest tutaj kluczowe, poprzez odniesienie do określonych własności środowiska naturalnego i społeczno-kulturowego. Stwierdziłem, że predykcje warunkują wybór pewnych działań, wykluczając tym samym inne. Są one więc ich naturalnym ograniczeniem, które można uznać za logiczne. Funkcja predykcyjna modelu jest oczywiście ograniczana, z jednej strony przez jego wewnętrzne parametry (struktura i treść sieci bayesowskiej oraz samej probabilistycznej reguły Bayesa), z drugiej zaś przez określone ograniczenia fizjologiczne (np. możliwości narządu wzroku; wydajność procesów neuronalnych) czy chemiczne (np. reakcje chemiczne zachodzące w komórkach piramidowych) itd.[47] Nie ulega jednak wątpliwości, że zależność pomiędzy przewidywaniem, że przykładowo spadnie deszcz, a wyborem taksówki jako środka lokomocji w przypadku oberwania chmury ma charakter logiczny[48]. Ową zależność logiczną określiłem wcześniej jako relację motywacyjną o postaci:

Jeżeli *A*, to *B*, *C* lub *D* (itd.), ale nie *E*, *F* lub *G* (itd.).

Relacja ta jest motywacyjna z powodu ograniczeń, jakie nakłada na wybór działań funkcja predykcyjna modelu, a ściślej rzecz ujmując – wygenerowana predykcja – oraz określone ograniczenia środowiskowe. Takie, a nie inne zachmurzenie stanowi określone ograniczenie fizyczne, wywodzące się z zachodzenia określonych zjawisk i procesów fizycznych w atmosferze. Oczywiście samo istnienie tego rodzaju ograniczenia nie warunkuje wyboru takiego, a nie innego działania. Jego rozpoznawalność jest ściśle powiązana z istnieniem zbioru odpowiednich przekonań, czyli z wierzytelnością modelu generatywnego, w który jest uzbrojony podmiot. Niemniej jest to realne ograniczenie, które ogranicza treść i charakter informacji docierającej do zmysłów. W zachmurzony dzień światło trafiające do siatkówki oka jest mocno ograniczone w stosunku do jego emisji w dzień bezchmurny.

Uwagi te uzasadniają wcześniejsze analizy dotyczące motywacji podmiotu. Relacja motywacyjna, która jest niejako ucieleśniona w relacji predykcji i określonych stanów świata, stanowi podstawę dla pojawienia się takiej, a nie innej motywacji bądź nastawienia u podmiotu. Zatem wyjaśnienie mechanizmów motywacyjnych powinno się odwoływać nie tylko do struktury modelu generatywnego i jego parametrów, jak sugeruje Friston i współpracownicy, ale także do już ustrukturyzowanego środowiska jako naturalnego ograniczenia dla podmiotu. Zatem sam wymóg maksymalizacji wiarygodności modelu nie może stanowić wystarczającego uzasadnienia dla takich, a nie innych działań bądź zachowań podmiotu. Co więc

[47] Nie oznacza to oczywiście, że chcąc wyjaśnić taki, a nie inny mechanizm zawsze musimy odwoływać się do niskopoziomowych ograniczeń powiązanych z biologią kory mózgowej czy fizjologią oka.

[48] Nie chodzi mi oczywiście o możliwość formalną, aczkolwiek i jej nie wykluczam, ale o możliwość materialną, którą można łatwo rozpoznać w tzw. wnioskowaniach materialnych (por. Brandom, 1994).

stanowi takie uzasadnienie? Istnienie pewnych określonych własności w środowisku, pośród których można wyróżnić m.in. afordancje.

5.4.2.3. Afordancje – wstępna analiza

W moich analizach wychodzę od rozumienia afordancji zaproponowanego przez Normana[49]. Twierdzi on, że afordancja to relacja, która wyznacza zakres wykorzystania danej rzeczy i odnosi się do związku pomiędzy określonymi własnościami przedmiotów a możliwościami podmiotu (może to być grupa społeczna, człowiek, ale też zwierzę, maszyna bądź robot). W tym ujęciu zależy ona od połączenia właściwości przedmiotu i możliwości podmiotu, który wchodzi w interakcję z tym przedmiotem. Afordancja nie jest zatem własnością, ale określoną relacją. Afordancja krzesła jako miejsca do siedzenia jest wypadkową pewnych własności tego przedmiotu (np. oparcie, określone wymiary fizyczne) oraz określonych predyspozycji podmiotu do wybierania takich, a nie innych przedmiotów do siedzenia (Norman, 2018, s. 29–31). Owe własności są, w proponowanym tutaj ujęciu, swoistymi ograniczeniami, które fundują relację afordancji. Oparcie krzesła i materiał, z którego jest wykonane, to fizyczne ograniczenia, które wykluczają możliwość siedzenia na nim przez więcej niż jedną osobę[50]. Podobnie określone własności podmiotu czynią z niego bycie podatnym do siedzenia na pewnego rodzaju krzesłach, jednocześnie ograniczając możliwość siedzenia na innych (np. krzesełku do karmienia w przypadku osoby dorosłej). Można zatem stwierdzić, że afordancje są ufundowane czy konstytuowane na podstawie określonych ograniczeń w środowisku. Afordancja bowiem z jednej strony oferuje wiele możliwości działania dla podmiotu ze względu na określone własności przedmiotu, z drugiej zaś szereg działań wyklucza jako niezasadne z punktu widzenia tych własności. Widząc krzesło, postrzegam je właśnie jako coś do siedzenia, ale już nie jako coś do pływania. Oczywiście w innych warunkach niż normalne (takich jak spędzanie czasu w domu) krzesło może mi się prezentować jako m.in. coś do wybicia szyby (np. gdy znajdę się w płonącym domu, a wybicie szyby pozwoli mi się z niego wydostać). Afordancje zatem mogą istnieć, twierdzi Norman, nawet wtedy, gdy nie są widoczne (Norman, 2018, s. 31). Nie da się ich bowiem zredukować do przekonań czy możliwości działania określonych podmiotów[51]. Uwagi te staną się jaśniejsze, gdy omówimy wcześniej przywołany przykład.

[49] Na temat pojęcia afordancji por. Chemero, 2003; Gibson, 1979.

[50] Oczywiście możemy sobie wyobrazić, że na krześle siada więcej niż jedna osoba (np. dwie), ale trudno sobie wyobrazić, aby na normalnym krześle siedziało dziesięć osób.

[51] Uwaga ta jest istotna dla dyskusji z tymi ujęciami, jakie zakładają możliwość wyjaśnienia działań poprzez odniesienie do wiedzy i przekonań podmiotów.

Gubię się w lesie i spotykam dzika. Możemy założyć, że postanowiłem nie wchodzić z nim w konfrontację, tylko uciec. Las oferuje wiele różnych możliwości działania. Oczywiście nie wszystkie są dla mnie widoczne. Królicza nora oferuje miejsce, do którego mogę wskoczyć. Afordancja ta jednak nie jest afordancją dla takiego podmiotu jak ja ze względu na szereg fizycznych ograniczeń. Przykładowo duże drzewo jest miejscem, na które mogę się wspiąć. Owa możliwość wspięcia się na nie, określana jest przez wiele różnych ograniczeń, takich jak gałęzie, kora, potencjalne zwierzęta. Owe ograniczenia konstytuują afordancję „czegoś do wspinania się" lub inaczej mówiąc: „czegoś, co pozwala się ukryć". Gałęzie stają się „czymś, po czym można się wspinać" itd. Inne afordancje oferuje otwarta przestrzeń, np. polana leśna. W niej mogę spróbować uciec przed dzikiem, o ile moje kondycyjne (fizyczne) ograniczenia mi na to pozwolą. Afordancja oferowana przez drzewo wiąże się z szeregiem różnego rodzaju ryzyk, ale teoretycznie oferuje miejsce, które jest bezpieczne. Afordancje polany leśnej otwierają przede mną możliwość pozostawienia dzika w tyle. Jakie działanie mam zatem wybrać? Mogę przewidzieć, że wspięcie się na drzewo szybko uczyni mnie bezpiecznym. Niemniej wiem, że mogę w trakcie tej czynności spaść i skręcić sobie kark albo złamać nogę. Przewiduję również, że bieg przez polanę, pomimo że prawdopodobnie nie grozi mi skręcenie karku, może zakończyć się uderzeniem przez dzika. Nie wiem bowiem, jak szybko dzik biegnie oraz czy moja kondycja pozwoli mi na wystarczające oddalenie się od niego.

Z punktu widzenia PP wybór pomiędzy działaniami opisanymi w przykładzie z dzikiem jest wyborem pomiędzy krótko- a długoterminową redukcją błędu predykcyjnego. W odniesieniu do fristonowskiego wnioskowania aktywnego powinno się wybrać działanie, które w większym stopniu zmaksymalizuje wiarygodność mojego modelu świata. Które jednak z tych działań mam wybrać? Zgodnie z regułą Bayesa wybieram takie przewidywanie, jakie z większym prawdopodobieństwem zminimalizuje błąd predykcyjny, odnoszący się w tym przypadku do wizji bolesnego kontaktu z dzikiem. Zgodnie z pierwszą predykcją błąd ten zminimalizuje działanie, jakim jest ukrycie się w konarach drzewa, zgodnie zaś z drugą dalsza ucieczka w las. Opisana sytuacja jest klasycznym przykładem działania w warunkach niepewności, czyli taką sytuacją, w jakiej nie znamy prawdopodobieństwa pojawienia się skutków wywołanych przez nasze działania. Przykład ten również jest radykalną wersją problemu ciemnego pokoju, ponieważ teoretycznie celem organizmu jest możliwie duża minimalizacja zaskoczenia. Problem polega jednak na tym, że wcale nie wiemy, które z podjętych przez nas działań w największym stopniu zminimalizuje owo zaskoczenie (wiąże się to bezpośrednio z istnieniem dywergencji K-L). Podobnie jest w ciemnym pokoju. Zupełnie nie wiemy, które z podjętych przez podmiot działań (czy zapalenie światła, czy opuszczenie pokoju, czy może pozostanie w nim) będzie najefektywniejsze. Twierdzę, że nie da się tego

rozstrzygnąć *a priori*. Każde działanie czy każda podjęta decyzja jest relatywna względem określonej sytuacji, warunków, nastawienia i motywacji.

Czy zatem powinniśmy pozostać sceptyczni w odniesieniu do rozwiązania przez PP problemu ciemnego pokoju i kompromisu eksploracyjno-eksploatacyjnego? Trudno na to pytanie dać jednoznaczną odpowiedź. Jedno jest jednak pewne: oba te problemy stanowią faktyczne ograniczenie dla tej ramy badawczej. Innymi słowy, być może samo PP nie jest wystarczające dla ich rozwiązania i potrzebne są albo dodatkowe założenia, albo inne metody badawcze, które takie rozwiązanie mogą zaoferować.

5.4.2.4. Granice wyjaśniania w przetwarzaniu predykcyjnym a postulat integracji

Wielu badaczy twierdzi, że FEP oferuje ramę eksplanacyjną, która pozwala zrozumieć, dlaczego wymóg minimalizacji błędów predykcyjnych znajduje swoje uzasadnienie w tym, że każdy żywy organizm dąży do homeostazy. W tym też sensie FEP miałaby stanowić teorię normatywną (por. podrozdział *Normatywność homeostazy*). Argumentacja zwolenników takiego rozwiązania opiera się na tym, że samo PP nie oferuje żadnej teorii decyzji czy np. algorytmów uczenia się. Biorąc pod uwagę taki argument, (1) można uznać, że moc eksplanacyjna PP odnosi się tylko do wycinka badań nad mózgiem czy niektórych badań nad percepcją (por. Sims, 2017); lub (2) stwierdzić, że PP powinno być oparte na jakiejś wiarygodnej teorii biologicznej, która osadza wymóg minimalizacji błędu predykcyjnego w szerszym kontekście; bądź (3) głosić, że PP podaje pewne ogólne ramy mechanistyczne lub mówiąc ściślej – szkic mechanizmu, który powinien zostać uzupełniony przez jakąś adekwatną teorię reprezentacji, procesów decyzyjnych, funkcji biologicznych itd. W niniejszej pracy w pewnym stopniu przyjmuję owo trzecie rozwiązanie, pokazując, że pełne wyjaśnienie procesów predykcyjnych zakłada przyjęcie nietrywialnej teorii mechanizmów normatywnych. Twierdzę dodatkowo, że nietrywialne i satysfakcjonujące wyjaśnienie procesów wyboru odpowiednich działań również zakłada przyjęcie „jakiejś" teorii podejmowania decyzji bądź określonego algorytmu tego mechanizmu. Wiąże się to m.in. z tym, co już wielokrotnie podkreślałem, że (1) pełne wyjaśnienie mechanizmów predykcyjnych musi wziąć pod uwagę zarówno wewnętrzne, jak i zewnętrzne ograniczenia. Jednak PP nie oferuje i nie może zaoferować dobrej teorii ograniczeń. A także z tym, że (2) jeżeli problem motywacji sprowadzi się, jak ma to miejsce w ramie aktywnego wnioskowania, do jednej naczelnej reguły działania, to problem ten nie zostanie satysfakcjonująco rozwiązany. Z tego też powodu twierdzę, że PP oferuje ogólną ramę badawczą, którą należy uzupełnić innymi ujęciami.

Uwaga ta koresponduje z przekonaniem, że pełne wyjaśnienie zjawisk poznawczych w kognitywistyce powinno się opierać, i faktycznie w wielu przypadkach tak jest, na mechanistycznej integracji poszczególnych teorii bądź modeli (por. Darden, Maull, 1977; Miłkowski, 2016a, 2016b; Mitchell, 2003; Povich, 2019). W tej propozycji chodzi o to, że nie poszukuje się jednej redukcjonistycznej teorii, która ma zunifikować wiedzę nad zjawiskami w danym obszarze badań, ile buduje się teorię na temat mechanizmów. Innymi słowy, poszukuje się tutaj wyjaśnień zjawisk w terminach mechanizmów, które je opisują. To, co wiąże ze sobą poszczególne modele bądź teorie, to wymóg wskazywania i opisywania mechanizmów. Przewaga tego podejścia nad innymi opiera się (1) na tym, że, zapewnia ona prosty sposób mówienia o poziomach; (2) oferuje znacznie więcej wglądu w to, czym jest integracja pomiędzy poziomami, na podstawie określonych ograniczeń, za pomocą których ocenia się relacje pomiędzy poziomami, a także koewolucję pracy na różnych poziomach, ich interakcje przyczynowo-skutkowe oraz organizację przestrzenną, czasową i hierarchiczną; oraz (3) opiera się na potrzebie wyjaśniania zjawisk zarówno oddolnie, jak i odgórnie, biorąc pod uwagę kontekst, w którym osadzone jest dane zjawisko. Integracja jest zatem próbą zobaczenia, w jaki sposób zjawiska na wielu różnych poziomach są ze sobą powiązane (Craver, Tabery, 2019)[52].

Carl F. Craver (2007) i Marcin Miłkowski (2016a, 2016b) wprost stwierdzają, że integracje w neuronaukach można opisać w terminach ograniczeń przestrzeni możliwych mechanizmów. Miłkowski dodaje jednak do tego, że niektóre zasady bądź reguły metodologiczne mogą kierować poszukiwaniami takich, a nie innych mechanizmów (są one również ograniczeniami) oraz że naprawdę satysfakcjonujące wyjaśnienie mechanizmów musi obejmować idealizację (Miłkowski, 2016b, s. 17). Takie ujęcie można wzbogacić przez perspektywę zaproponowaną przez Danksa (2014, s. 31): „określona teoria *S* ogranicza inną teorię *T*, jeśli stopień, w jakim *S* ma jakąś teoretyczną cnotę *V* (np. prawda, moc predykcyjna, moc wyjaśniająca), ma znaczenie w zakresie, w jakim *T* ma *V*". Innymi słowy, jeśli *S* ogranicza *T* z powodu pewnej cnoty teoretycznej, to jeżeli przywiązujemy wagę do tej cnoty w *T*, to powinniśmy przywiązywać do niej wagę również w *S*. W tym ujęciu cnota oznacza określone ograniczenie.

5.4.2.5. Ograniczenia jako warunki integracji modeli

W proponowanym tutaj modelu mechanizmów normatywnych kluczowe znaczenie spełniają ograniczenia. Są one bowiem tym elementem, który czyni dany mechanizm nie tylko przyczynowym, ale także normatywnym w sensie przeze

[52] Jako historyczne przykłady tego podejścia można podać odkrycie syntezy białek (Darden, 2006) bądź biologii komórki (Bechtel, 2006).

mnie podanym. Stwierdziłem wcześniej, że kluczowe dla tego typu mechanizmów są określone stany środowiska (fizycznego, społeczno-kulturowego itd.), które można analizować w kategoriach ograniczeń dla mechanizmów. Owe ograniczenia nie pełnią tylko funkcji kontekstu, w którym mechanizm jest realizowany, ale przede wszystkim są jego konstytutywnym komponentem. Przykładowo wyjaśnienie takiej własności predykcji generowanych przez mechanizm predykcyjny, jaką jest weredyczność, nie jest możliwe, jeżeli nie weźmie się pod uwagę faktycznych warunków jej stwierdzalności (Burge, 2010). Twierdzę, że pełne wyjaśnienie mechanizmów normatywnych, analizowanych tutaj na przykładzie mechanizmów predykcyjnych, zakłada określoną integrację poszczególnych modeli takich, a nie innych zjawisk poznawczych. Należy zatem powiedzieć, że mechanizmy normatywne stanowią ograniczenie, które sprawia, że PP powinno być zintegrowane z takim modelem (lub modelami), które umożliwiają pełne ich wyjaśnienie. Zasadność tego stwierdzenia potwierdza nasza teza dotycząca słabości eksplanacyjnej tych ujęć, które sprowadzają wyjaśnienie wybranych zjawisk poznawczych do jednego, prostego wymogu (np. długoterminowej minimalizacji błędu predykcyjnego), jednocześnie przyjmując monizm eksplanacyjny. Innymi słowy, jeżeli mechanizmy predykcyjne bądź reprezentacyjne są mechanizmami normatywnymi, a twierdzę, że są, to ich wyjaśnienie w aspekcie realizowanych przez nie funkcji normatywnych wykracza poza ramy badawcze właściwe dla PP z tego powodu, że rama ta nie oferuje satysfakcjonującej i nietrywialnej koncepcji ograniczeń dla tychże mechanizmów. Dopiero (mechanistyczna) integracja PP (bądź ramy aktywnego wnioskowania) z innym modelem, *resp.* modelami, pozwoli na dobre wyjaśnienie natury normatywnych mechanizmów predykcyjnych. Twierdzę również, że niemożliwość rozwiązania problemu ciemnego pokoju przez obie te ramy jest bezpośrednim następstwem monizmu eksplanacyjnego przyjętego przez zwolenników PP[53]. Pokazałem to na przykładzie problemu motywacji.

Istnienie takich, a nie innych ograniczeń dla określonych zjawisk jest powodem, dla którego powinno się integrować różne modele, pozwalające na rekonstrukcję wybranych mechanizmów. Zdaniem Dardena (2006, s. 272) proces ich rekonstrukcji polega na konstruowaniu, ocenie i rewizji schematów określonych mechanizmów w odniesieniu do empirycznych i eksperymentalnych ograniczeń. Innymi słowy, ograniczenia zawężają przestrzeń możliwych mechanizmów do określonego obszaru, który ma zrekonstruować model. Z tego powodu integracja PP z innymi modelami pozwoli na satysfakcjonującą rekonstrukcję mechanizmów normatywnych, do których należy zaliczyć mechanizmy predykcyjne. PP oferuje taki, a nie inny model przetwarzania informacji, obiecując przy tym dobre wyjaśnienie sprzężenia procesów percepcyjnych z procesami kontroli działania czy podejmowania decyzji. Ta pretensja eksplanacyjna nie jest, jak twierdzi wielu

[53] O ograniczeniach eksplanacyjnych PP w odniesieniu do problemu myślenia pisze m.in. Williams, 2018.

przeciwników tego ujęcia, przedwczesna czy nieuzasadniona, a wręcz przeciwnie: twierdzę, że stanowi ona wyraz tego, że PP należy zintegrować z innymi modelami zjawisk poznawczych. Wbrew literalnym wypowiedziom wielu zwolenników tego ujęcia faktyczna praktyka badawcza pokazuje, że PP wielokrotnie wykorzystuje na zasadzie heurystycznej bądź metodologicznej rozwiązania przyjęte w innych modelach[54]. W niniejszych rozważaniach bronię przekonania, zgodnie z którym dopiero podejście pluralistyczne do wyjaśniania pozwala w pełni wykorzystać możliwości zawarte w PP.

5.5. Wyjaśnienie mechanizmów normatywnych w kontekście przetwarzania predykcyjnego

W poprzednim podrozdziale ustaliliśmy, że pełne wyjaśnienie mechanizmów normatywnych oraz ich specjalnego przypadku, jakim są mechanizmy predykcyjne, zakłada uwzględnienie istotnego komponentu, którym są stany środowiska. Zaproponowano, aby stany środowiska analizować w terminach ograniczeń dla mechanizmów. Ograniczenia mogą mieć charakter fizyczny, kulturowy, logiczny, ale także metodologiczny w sensie bycia normą bądź zasadą określającą sposób wyjaśniania (Miłkowski, 2016a, s. 50). Moc eksplanacyjna ograniczeń w wyjaśnieniach jest uwarunkowana faktyczną rolą przyczynową, jaką ograniczenia grają w takich, a nie innych zjawiskach, które chcemy wyjaśnić. Oznacza to, że wyjaśnia się określone zjawisko poprzez odniesienie do komponentu, jakim jest takie, a nie inne ograniczenie, ponieważ sam mechanizm odpowiedzialny za realizację tego zjawiska zawiera taki, a nie inny komponent. Takie wyjaśnienie w przypadku badanych tutaj normatywnych mechanizmów predykcyjnych zakłada z konieczności integrację z innymi modelami, które mogą zaoferować satysfakcjonujące wyjaśnienie ograniczeń dla tychże mechanizmów. Analiza tych modeli znacznie wykracza poza ramy przyjęte dla niniejszych dociekań. Niemniej nie wyklucza to możliwości doprecyzowania postulowanego tutaj schematu mechanizmu normatywnego. Twierdzę, że zaletą tego schematu jest możliwość wyjaśnienia wielu zjawisk powiązanych z funkcjonowaniem podmiotów uzbrojonych w probabilistyczne modele świata. Jeżeli PP ma być zasadną ramą wyjaśniającą sprzężenie percepcji, działania i wyższych procesów poznawczych, za czym argumentowałem we wcześniejszych rozdziałach, to analiza mechanizmów normatywnych pozwala w sposób pełniejszy wyjaśnić owo sprzężenie, jednocześnie wyraźnie pokazując ograniczenia tego podejścia.

[54] Dobrym przykładem jest książka Clarka *Surfing uncertainty*, w której wielokrotnie odwołuje się on do określonych modeli pamięci, percepcji bądź kategoryzacji.

Istnienie mechanizmów normatywnych jednoznacznie dowodzi, że istnieją zjawiska, których wyjaśnienie wykracza poza standardowe wykładnie PP i FEP. Trudność z rozwiązaniem problemu ciemnego pokoju oraz kompromisu eksploracja – eksploatacja bezpośrednio wiąże się z przyjmowanym przez wielu badaczy monizmem eksplanacyjnym, który w tym przypadku opiera się na przekonaniu, że PP jest wystarczającą ramą eksplanacyjną dla wyjaśnienia wielu zjawisk. Jesteśmy zgodni z tym, że bayesizm w wersji przyjętej w PP pozwala w nietrywialny sposób wyjaśnić szereg zjawisk, jednak jego moc eksplanacyjna jest w pewnym sensie mierzona możliwością zintegrowania go z innymi modelami (por. Williams, 2018). W innym przypadku jesteśmy zmuszeni bronić jakiejś potencjalnej wersji teorii normatywnej, której zasadność została przez wielu badaczy podważona (por. Elqayam, Evans, 2011; Jones, Love, 2011; Kahneman, Tversky, 1979; Kahneman, 2012; także podrozdział *Normatywność bayesowskiego modelu przetwarzania predykcyjnego*).

Wciąż zatem bronię przyjętej w tej pracy hipotezy ontycznej, która głosi, że istnieją takie zjawiska, jakie wyjaśnia się przy pomocy bayesowskiego PP, ponieważ są one bayesowskie z natury (a jeżeli są one bayesowskie z natury, to znaczy, że powinno się je wyjaśniać przy użyciu modelowania probabilistycznego). Niemniej należy tę hipotezę uzupełnić o dodatkowe założenie, które brzmi następująco: tylko takie modele bayesowskie są nietrywialne eksplanacyjnie, jakie umożliwiają wyjaśnienie swoich eksplanandów w odniesieniu do wszelkich konstytutywnych ograniczeń[55]. W praktyce badawczej oznacza to, że nietrywialne modele Bayesa muszą dać się zintegrować z innymi modelami bądź teoriami. Takim modelem jest PP. Analogicznie należy uzasadnić normatywny charakter tego modelu. Model ten nie jest normatywny w takim sensie, że formułuje jedną bądź więcej zasad normatywnych, ale dlatego, że wyjaśnia określone zjawiska poprzez rekonstrukcję i opis mechanizmów normatywnych, które są odpowiedzialne za ich realizację. Innymi słowy, bayesowskie PP jest normatywne, ponieważ (przynajmniej niektóre) zjawiska przezeń wyjaśniane są normatywne z natury. Takie ujęcie PP umożliwia wybronienie go przed zarzutami, jakie formułują zwolennicy teorii empirycznych wobec koncepcji preskryptywnych. Normatywność PP wywodzi się z normatywności mechanizmów, które spełniają określone funkcje przyczynowe. Nie ma tutaj mowy o żadnych regułach bądź normach, które stanowić miałyby swoistą miarę poprawności, skuteczności bądź optymalności określonych działań i zachowań.

[55] Nie wszystkie bowiem ograniczenia są istotne dla wyjaśnienia określonych mechanizmów. Część z nich może stanowić tylko tło, inne zaś mogą faktycznie grać takie, a nie inne role przyczynowe.

5.5.1. Problem ciemnego pokoju a mechanizmy normatywne

Kluczową kwestią dla rozwiązania problemu ciemnego pokoju jest możliwość wyjaśnienia motywacji. Po przeprowadzeniu dokładniejszych analiz należy zgodzić się z Kleinem, że ani predykcje, ani minimalizacja energii swobodnej nie wystarczą, aby wyjaśnić adaptację organizmów oraz ich motywację (Klein, 2018, s. 4). Zdaniem Kleina, zarówno w przypadku PP, jak i ramy aktywnego wnioskowania opartej na FEP, mamy do czynienia z daleko posuniętą idealizacją rozumienia systemów poznawczych, zgodnie z którą organizmy są optymalnymi regulatorami w sensie Conanta i Ashby'ego (por. Conant, Ashby, 1970), ponieważ nieustannie popełniają one błędy[56]. Nie da się sprowadzić wszystkich mechanizmów regulacyjnych, do których zgodnie z PP i FEP można zaliczyć motywację, do zmiany wewnętrznych parametrów modelu (Klein, 2018, s. 16). Z tego też powodu, przykładowo, wielu pilotów bądź chirurgów zwiększa swoje umiejętności, minimalizując zaskoczenie, w ten sposób, że przenoszą szereg działań regulujących na listy kontrolne (por. Gawande, 2010). Remedium na problemy PP i FEP z motywacją ma być, zdaniem Kleina, rezygnacja z części ambicji eksplanacyjnych. Moim zdaniem Klein odsłania tylko część problemu. Główną trudnością tych ram badawczych nie są tylko wygórowane ambicje badawcze, ale przede wszystkim zakładany *implicite* monizm eksplanacyjny. Oznacza to, że rozwiązanie problemu ciemnego pokoju przekracza możliwości eksplanacyjne PP.

Jak zatem rozwiązać ten problem? Twierdzę, że należy uwzględnić również wiele ograniczeń, które kształtują motywację (por. podrozdział *Znaczenie ograniczeń dla działania*). To właśnie ustrukturyzowane przez ograniczenia środowisko, w sposób konstytutywny, a nie tylko przyczynowy, wpływa na motywację, której celem jest, przypomnijmy, minimalizacja niepewności. Brak uwzględnienia roli komponentów środowiskowych prowadzi do niemożliwości wyjaśnienia motywacji i do regresu w nieskończoność (por. podrozdział *Uwagi wobec ramy aktywnego wnioskowania*): motywację do poszukiwania nowości uzasadnia się poprzez odniesienie do takich zasad działania (*policy*), które kształtują taką, a nie inną potrzebę motywacyjną. Co zatem uzasadnia takie zasady? Zapewne jakieś wysokopoziomowe przekonania? A jakie jest źródło tych przekonań? I tak dalej.

Przykłady pilotów i lekarzy, o których wspomina Klein, dobitnie unaoczniają potrzebę uwzględnienia zewnętrznych ograniczeń w analizie mechanizmów minimalizujących niepewność. Piloci niejako oddelegowują część swoich zadań na listy kontrolne, aby móc poprawić wydajność swoich własnych działań, tym samym

[56] Por. „Zgodnie z zasadą energii swobodnej, podmiot staje się optymalnym (w przybliżeniu) modelem swojego środowiska. Jest tak dlatego, że matematycznie zaskoczenie jest również negatywnym dowodem logarytmicznym dla modelu podmiotu. Oznacza to, że minimalizacja zaskoczenia maksymalizuje dowody dla modelu. Mówiąc wprost, podmiot staje się modelem środowiska, w którym jest osadzony. Jest to dokładnie zgodne z twierdzeniem Conanta i Ashby'ego o dobrym regulatorze" (Friston i in., 2012b, s. 2).

minimalizując niepewność. Owe listy bądź wykazy kontrolne przygotowuje się dla trudnych zadań w celu (1) zapewnienia im właściwej, tj. optymalnej kolejności, w taki sposób, aby (2) żaden istotny etap postępowania nie został pominięty[57]. W tym przykładzie lista kontrolna jest istotnym ograniczeniem, które umożliwia lepsze wykonanie zadania postawionego przed pilotami bądź chirurgami, które w odniesieniu do PP wiązać należy z długoterminową minimalizacją uśrednionego błędu predykcyjnego (Hohwy, 2013, s. 85). Lista kontrolna stanowi tutaj ograniczenie, które kształtuje motywację w tym sensie, że przyczynia się do zminimalizowania niepewności, czyli zmienia rozkład prawdopodobieństwa przestrzeni, w której realizowany jest mechanizm pilotowania samolotu. Lista kontrolna jest tutaj rozumiana nie jako ograniczenie fizyczne, bowiem jej nośnik nie jest istotny z punktu widzenia realizacji mechanizmu, ale jako ograniczenie logiczne i konwencjonalne (metodologiczne). Z jednej bowiem strony wskazuje ona kolejność doboru odpowiednich działań, z drugiej zaś określa, jakie dokładnie działania metodycznie powinny zostać podjęte, aby osiągnąć sukces.

Jak zatem wyjaśnić zjawisko, jakim jest pilotowanie samolotu? Jest to oczywiście złożony mechanizm, który można sprowadzić do szeregu bardziej podstawowych. Owa redukcja do bardziej podstawowych mechanizmów nie jest krokiem koniecznym, bowiem to strategia eksplanacyjna i zadania badawcze decydują o tym, na jakim poziomie abstrakcji i organizacji zjawiska chcemy wyjaśniać. Należy powiedzieć, że jednym z mechanizmów fundujących omawiane tutaj działanie jest mechanizm predykcyjny odpowiedzialny za realizację zadania, jakim jest szacowanie ryzyka w warunkach niepewności. Pilot bowiem musi możliwie skutecznie unikać wielu zagrożeń, które wiążą się z pilotażem samolotu (np. możliwości dezorientacji przestrzennej, pomylenia procedur, niewłaściwego odczytania narządów pomiarowych). Skuteczna minimalizacja tego rodzaju ryzyka wiąże się z koniecznością wykorzystania m.in. listy kontrolnej[58]. Zatem wyjaśnienie minimalizacji błędów predykcyjnych powinno w tym kontekście odwołać się (1) do takich komponentów mechanizmu predykcyjnego jak wygenerowana predykcja, określone parametry modelu generatywnego oraz właśnie komponent ograniczeń (tj. lista kontrolna), (2) struktury mechanizmu, czyli wewnętrznej architektury modelu hierarchicznego, oraz (3) operacji, które on realizuje. Są to m.in. utrzymywanie stabilności systemu, odnoszenie się do określonych stanów środowiska, na które jest skierowana predykcja, oraz ocena wartości epistemicznej treści wygenerowanej predykcji (jako mniej lub bardziej prawdopodobnej w sensie Bayesa) (por. Miłkowski, 2015a, s. 712). Mechanizm predykcyjny wyjaśnia swoje zjawisko tylko wtedy, gdy jego komponenty odgrywają określone role przyczynowe. Oznacza to, że można o nich

[57] Por. Hutchins, 1995, na temat analizy nawigacji amerykańskiego statku wojennego jako przykładu rozproszonego procesu poznawczego.

[58] Kluczowa jest również np. współpraca pomiędzy pilotami.

orzekać, że są przyczynami określonych działań podmiotu zrelatywizowanych do warunków otoczenia. Innymi słowy, wskazanie na normatywny mechanizm predykcyjny pozwala na wyjaśnienie, dlaczego takie, a nie inne działanie pilota wykorzystującego listę kontrolną może być skuteczne. Samo zachodzenie mechanizmu nie gwarantuje, że będzie ono skuteczne, ale określa warunki, w których może ono takie być. Przyczyną wykorzystania przez pilota m.in. listy kontrolnej jest predykcja, zgodnie z którą wykorzystanie tego typu artefaktu pozwoli na skuteczniejszą minimalizację potencjalnych błędów predykcyjnych. Jednak pełne wyjaśnienie tego działania sformułuje się dopiero wtedy, gdy uwzględni się w analizie mechanizmu przyczynową rolę zewnętrznego ograniczenia, jakim jest w tym przypadku lista kontrolna. Mechanizm jest normatywny, ponieważ odgrywa on określone role przyczynowe, wskazując tylko te przyczyny, które są dla danego zjawiska konstytutywne, oraz określa warunki środowiskowe, które ograniczają możliwą przestrzeń dla zachodzenia danego zjawiska.

W tym kontekście można wyjaśnić możliwość pojawienia się motywacji. W rozdziale *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego* zdefiniowałem relację motywacyjną (na której jest ufundowana motywacja jako potrzeba redukcji niepewności) jako zależność pomiędzy predykcjami a kierowanymi przez nie działaniami w odniesieniu do określonych stanów środowiska (por. podrozdział *Przetwarzanie predykcyjne a mechanistyczny model wyjaśniania naukowego*). Oznacza to, że motywacja podmiotu jest kształtowana przez wygenerowane predykcje, które z jednej strony odnoszą się do konieczności minimalizacji potencjalnych błędów predykcyjnych, przy uwzględnieniu stanów środowiska, z drugiej zaś przez możliwości działań, które są oferowane przez to środowisko. To, co jest tutaj istotne, to stwierdzenie, że stanów środowiska nie można sprowadzić do funkcji bycia nagrodą, karą bądź dowodem zwiększającym wiarygodność modelu. Lista kontrolna jest istotna dla wyjaśnienia działania pilotów, ale nie dlatego, że stanowi ona potencjalną nagrodę lub wartość dla pilota. Również nie dlatego, że w pewnym sensie zwiększa wiarygodność jego modelu świata, ale dlatego, że spełnia ona określone funkcje przyczynowe w wyjaśnieniu działań, które pilot podejmuje. Stwierdzenie, że maksymalizuje ona wiarygodność takich, a nie innych przekonań, zwiększając tym samym skuteczność określonych działań podmiotu, nie jest eksplanacyjnie istotne. Eksplanacyjnie istotne jest to, że bez odniesienia do takiego ograniczenia, jakim jest lista kontrolna, nie da się wyjaśnić mechanizmu predykcyjnego, który warunkuje skuteczne pilotowanie samolotu[59].

Przykład działania pilotów trafnie unaocznia problem ciemnego pokoju w tym sensie, że wskazuje na dynamikę wyjaśnień budowanych w odniesieniu do kon-

[59] Przykład listy kontrolnej dobitnie pokazuje, że pełne wyjaśnienie wielu zjawisk pociąga za sobą konieczność uwzględnienia wielu perspektyw badawczych. PP bowiem nie jest teorią, która może wyjaśnić skuteczność działania takich artefaktów czy w szerszej perspektywie – tego typu ograniczeń.

tekstu środowiskowego. Oznacza to, że nie da się przy pomocy jednej reguły czy wymogu normatywnego powiedzieć, że podmiot w takich, a nie innych warunkach powinien tak a tak się zachować, że np. zawsze zapali światło, albo że będzie chciał opuścić zaciemnione pomieszczenie, albo że będzie chciał w nim pozostać. To, jak się podmiot zachowa, jest regulowane bowiem nie tylko przez parametry wewnętrznego modelu świata, biologiczny wymóg minimalizacji niepewności, ale także przez jego motywację, która jest warunkowana przez niezależne od niego środowisko. Pełne wyjaśnienie możliwości motywacji podmiotu wykracza poza możliwości eksplanacyjne PP i FEP. Należy bowiem wziąć tutaj pod uwagę *jakieś* nietrywialne teorie ograniczeń, preferencji czy emocji. Twierdzę jednak, że działania podmiotów można wyjaśnić poprzez wskazanie leżących u ich podstaw mechanizmów normatywnych. Jednak ich pełny opis zakłada integrację wielu modeli i podejść badawczych[60].

5.5.2. Rola predykcji

Jak dotąd nie została należycie omówiona funkcja samych predykcji w mechanizmach predykcyjnych. W rozdziale *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego*, omawiając ogólny schemat tzw. relacji motywacyjnej[61] konstytuującej zależność pomiędzy przewidywaniem a działaniem (por. podrozdział *Przetwarzanie predykcyjne a mechanistyczny model wyjaśniania naukowego*), stwierdziłem, że predykcje warunkują pojawienie się określonych działań, wykluczając tym samym zajście innych. Wybór takiego, a nie innego działania zależy od wygenerowanej predykcji. Proces selekcji predykcji opiera się na (1) bayesowskim wnioskowaniu abdukcyjnym; (2) zrelatywizowanym do takiej, a nie innej strategii działania przyjętej przez dany podmiot; (3) podstawie jego regulowanej wewnętrznie (parametry modelu generatywnego i wysokopoziomowe przekonania) i zewnętrznie (określone ograniczenia środowiskowe) motywacji. Oznacza to, że predykcja warunkująca działanie jest wybierana na podstawie pewnego potencjału zaimplementowanej w modelu generatywnym sieci bayesowskiej oraz określonych relacji ze środowiskiem. W praktyce oznacza to, że podmiot wybiera wśród kilku kontrfaktycznych hipotez o postaci „co się stanie, jeżeli zrobię to a to" (Piekarski, 2019, s. 2; por. także Seth, 2015). Modelowanie probabilistycznych scenariuszy działań pozwala na planowanie działań i długoterminową minimalizację błędów predykcyjnych (Pezzulo, Rigoli, Friston, 2015, s. 24; por. także Clark, 2019, s. 10). To, co jednak odróżnia proponowane tutaj ujęcie od innych obecnych w literaturze, to podkreślenie eksplanacyjnej roli czynników środowiskowych. Działanie

[60] Postulat integracji PP z innymi podejściami sformułowali m.in. Williams (2018) i Klein (2018).
[61] Jego podstawowa forma jest następująca: „Jeżeli '*predykcja*' (warunek), to '*działanie*' (skutek)".

bowiem jest skuteczne, jeżeli wygenerowana przez model predykcja uwzględnia takie, a nie inne stany otoczenia. Oznacza to, że precyzyjna (tj. wysoko ważona) (kontrfaktyczna) predykcja musi korespondować z określonymi faktycznymi bądź kontrfaktycznymi własnościami świata[62]. W ten sposób wysokopoziomowe precyzyjnie ważone predykcje określają sposób, w jaki działamy w świecie.

Trzeba zatem powiedzieć, co już parokrotnie podkreślałem, że podejmowane przez podmiot działania są normowane nie tylko przez takie, a nie inne predykcje, ale również przez określone stany czy własności otoczenia, które są rozumiane w niniejszej pracy jako ograniczenia dla mechanizmu. Dzięki temu predykcje nie tylko kierują działaniami, ale także kształtują przyczynowe przejścia pomiędzy stanami, które mają określoną treść i warunki spełniania (np. stany mentalne). Bronione tutaj stanowisko należy zatem określić mianem „eksternalistycznego", przez co rozumiem, że egzemplifikacja określonych stanów mentalnych jest warunkowana nie tylko przez parametry modelu generatywnego (np. określone pobudzenia populacji neuronalnych w przypadku stanów niskopoziomowych), ale także przez takie, a nie inne stany środowiska i społeczno-kulturowe usytuowanie podmiotów. To zatem normatywna relacja predykcji ze środowiskiem determinuje bądź specyfikuje treść określonych stanów. Dlaczego normatywna, a nie po prostu przyczynowa? Bowiem tej relacji, która jest konstytutywna dla interakcji podmiotu z otoczeniem, nie da się, co wykazałem, sprowadzić do struktury modelu określonych algorytmów uczenia się czy wymogów minimalizacji błędów predykcyjnych bądź maksymalizacji wiarygodności tego modelu. Dodatkowo umożliwia ona wyjaśnienie znaczenia błędów poznawczych (np. błędu reprezentacyjnego) dla organizmu z jego własnej perspektywy, a nie tak jak ma to miejsce np. w teleosemantyce, czyli z perspektywy zewnętrznego obserwatora. Oznacza to, że możliwy błąd lub błędna reprezentacja ma normatywne znaczenie dla organizmu, tj. wpływa na dobór i kierowanie działaniami, a nie tylko jest potencjalnym efektem określonych procesów przyczynowych.

Przeprowadzone tutaj analizy prowadzą do wzmocnienia tezy dotyczącej normatywności mechanizmów predykcyjnych. Relacja motywacyjna zachodząca pomiędzy wyselekcjonowaną predykcją a działaniem bądź ich ciągiem, podjętych ze względu na ową predykcję, może zostać utrwalona w czasie. Mianowicie: statystyczna skuteczność działań podejmowanych w takich, a nie innych warunkach może prowadzić do wytworzenia takiego, a nie innego wzorca zachowań. Chodzi o to, że jeżeli w określonych okolicznościach pewna predykcja warunkująca określone działanie bądź działania prowadzi do zamierzonych przez podmiot celów, to pod-

[62] Godfrey-Smith twierdzi, że istnienie korespondencji pomiędzy stanami podmiotu a określonymi stanami świata jest gwarantem skuteczności działań (por. Godfrey-Smith, 1996, rozdział 6). W PP nietrywialna koncepcja wyjaśniająca korespondencję określonych parametrów modelu generatywnego i stanów świata wykorzystuje tzw. reprezentacje strukturalne (S-reprezentacje) (por. Gładziejewski, 2016; Kiefer, Hohwy, 2018; a także podrozdział *S-reprezentacje i realne wzorce*).

miot może wyuczyć się pewnego schematu działania. Schemat ten można nazwać „wzorcem zachowania" w tym sensie, że stanowi on pewną matrycę określającą, jak należy się zachować w takich, a nie innych warunkach środowiskowych (por. Churchland, 2002)[63]. Statystyczna skuteczność może zatem prowadzić do wytworzenia pewnych reguł działania, które można sprowadzić do następującej formy:

> Jeżeli w warunkach *X*, optymalnym działaniem jest *S*,
> to jeżeli podmiot *A* chce działać możliwie optymalnie,
> to *A* w *X* powinien wykonać *S*[64].

Wyjaśnieniem działania *S* jest, zgodnie z bronioną tutaj koncepcją mechanizmów normatywnych, wskazanie jego konstytutywnej przyczyny na poziomie mechanizmu, czyli określonej predykcji wygenerowanej przez mechanizm predykcyjny. Uzasadnieniem zaś takiego działania jest podana powyżej reguła. Takie uzasadnienie może przybrać następującą postać: Zrobiłem *S*, ponieważ w warunkach *X* takie działanie było optymalne. W tym sensie wskazanie reguły jest jednocześnie podaniem racji przemawiającej za takim, a nie innym działaniem podmiotu[65].

5.6. Mechanizmy normatywne a hipoteza rozproszonej normatywności

Wiemy już, że pełne wyjaśnienie mechanizmów normatywnych zakłada uwzględnienie zarówno czynników wewnętrznych (określone parametry modelu generatywnego i stany podmiotu), jak i zewnętrznych: stany środowiska, ograniczenia, afordancje, czyli wszystko to, co się składa na naturalne, jak i materialno-społeczno-kulturowe usytuowanie podmiotu. Nie można zatem normatywności na tym poziomie sprowadzić do jakiegoś jednego elementu, raczej należy ją wiązać z wieloma różnymi. Oznacza to, że analizowane tutaj mechanizmy normatywne stanowią fragment większej całości. Przypomnijmy: o normatywnym charakterze określonego mechanizmu decydują nie tylko jego funkcje odniesione do organizacji i struktury organizmu, do którego stabilności dany mechanizm się przyczynia, ale także odniesienie do określonych stanów środowiska, które jako komponenty mechanizmu są konstytutywne dla wyjaśnienia danego zjawiska. Trafnie owo odniesienie do stanów środowiska określa Millikan, mówiąc o strukturach rela-

[63] Owo „należy" trzeba rozumieć w sensie skuteczności pragmatycznej (tzn. że lepiej jest wybrać taki wzorzec działania w tych okolicznościach niż inny), a nie w sensie powinności moralnej.

[64] Tak sformułowane reguły można przekształcić w normy mające charakter bezwarunkowy: „W takich, a nie innych warunkach powinno się robić to, a to".

[65] Uwagi te pozwalają na sformułowanie wstępnych hipotez dotyczących zastosowania PP do wyjaśniania zjawisk na poziomie osobowym (por. na ten temat Clark, 2017b, 2019).

cyjnych (Millikan, 2011, s. 122–123). Twierdzę, że to właśnie owe struktury mają charakter normatywny. Łatwo to wyjaśnić na przykładzie normatywnych mechanizmów predykcyjnych. Jeżeli analizuje się je jedynie z perspektywy podmiotu, *resp*. modelu generatywnego, to mechanizmy predykcyjne bądź reprezentacyjne są jednym z rodzajów mechanizmów przyczynowych. Dlatego też niektórzy badacze (np. Gładziejewski, 2015, 2016; Gładziejewski, Miłkowski, 2017) twierdzą, że wystarczy powołać się na same reprezentacje, aby wyjaśnić działania podmiotu, tzn. że określone własności reprezentacji czynią z nich dobre i skuteczne przewodniki działań. Takie ujęcia jednak nie wyjaśniają mechanizmu normatywnego, czyli konstytutywnego przyczynowo, który warunkuje wybór takiego, a nie innego działania[66]. One określają jedynie ich przyczynę w kategoriach gatunkowych. Wyjaśnienia oparte na mechanizmach normatywnych pozwalają nie tylko wskazać przyczynę gatunkową, ale również pozwalają dać odpowiedź na pytanie o to, dlaczego dany podmiot podjął w takich okolicznościach właśnie te, a nie inne działania.

Jeżeli weźmie się pod uwagę dodatkowo aspekt środowiskowy mechanizmów normatywnych, to okazuje się, że stanowią one element pewnej szerszej struktury normatywnej, na którą składają się zarówno określone ograniczenia, z których przynajmniej część ma charakter normatywny [por. lista zakupów u Marra (1982)], rzeczywiste wzorce (por. Churchland, 2002; Dennett, 2008), afordancje rozumiane nie tylko jako możliwości działań oferowane przez środowisko, ale także jako określone normy działania (por. Heras-Escribano, de Pinedo, 2016; Piekarski, Wachowski, 2018; Ramstead, Veissière, Kirmayer, 2016; Wachowski, 2018a); w dalszej kolejności normy i konwencje społeczne itd. Jeżeli zatem chcemy wyjaśnić w pełni mechanizmy normatywne, to należy wziąć pod uwagę również inne elementy tej szerszej struktury normatywnej i zastanowić się nad tym, które z tych elementów są faktycznie konstytutywne dla badanych tutaj mechanizmów, a które stanowią tylko element tła. Twierdzę zatem, że należy przyjąć, że wyjaśnienie mechanizmów normatywnych nie powinno odwoływać się jedynie do określonych elementów owej struktury, ale raczej do relacji, w jakie te elementy wchodzą, konstytuując tym samym określone komponenty, operacje i dalsze struktury. Z tego też powodu przyjmuję w tej pracy hipotezę czegoś, co od tej chwili będę określał mianem „normatywności rozproszonej". Hipoteza normatywności rozproszonej zakłada, że o normatywnym charakterze określonych mechanizmów, funkcji czy przedmiotów decydują relacje, w które te mechanizmy, funkcje czy przedmioty wchodzą. Przykładowo: rozpoznanie normatywnego charakteru funkcji predykcyjnej zasadza się na odniesieniu jej do działań, które są przez nią warunkowane w odniesieniu do określonych stanów środowiska.

[66] W niniejszej pracy mechanizm reprezentacyjny również może być określonym mechanizmem normatywnym.

Można by zatem powiedzieć, że normatywne są przede wszystkim określone relacje, a nie przedmioty, które stanowią elementy czy komponenty tych relacji. Z tego też powodu mechanizmy predykcyjne można określić jako przyczynowe mechanizmy normatywne, zaś samą normatywność jako konstytutywną[67]. Jest tak, bowiem spełniają one określone konstytutywne funkcje przyczynowe dla podejmowanych decyzji i realizowanych przez podmiot działań. Innymi słowy, mechanizm predykcyjny jest normatywny, bowiem konstytuuje on określoną relację motywacyjną (pomiędzy predykcją a działaniem), której nie można sprowadzić do prostej relacji przyczynowej o formie „Jeżeli *A*, to *B*" (por. podrozdział *Przetwarzanie predykcyjne a mechanistyczny model wyjaśniania naukowego*).

Uzasadnienie hipotezy normatywności rozproszonej wymaga dodatkowych badań, które wykraczają poza cele wyznaczone dla tej pracy. Niemniej przedstawione tutaj dociekania pozwalają na wstępne uznanie prawdziwości tej hipotezy. Jeżeli bowiem chce się dokonać analizy mechanizmów predykcyjnych w aspekcie ich normatywności, to konieczne jest odwołanie się do tych komponentów, które wykraczają poza indywidualne uwarunkowania organizmu bądź podmiotu. Należy wskazać te elementy środowiska, które współkonstytuują dany mechanizm.

Określone własności środowiska, ograniczenia i afordancje tworzą w pewnym sensie złożone konstelacje, szersze normatywne struktury. Dobrym przykładem takiej szerokiej czy wręcz rozproszonej struktury normatywnej jest miasto albo w mniejszej skali ulica. Tutaj bowiem wiele elementów ma określone znaczenie, wyznaczając tym samym szereg reguł działania (np. przepisy drogowe, sygnalizacja świetlna, znaki na drodze). Pasy stanowią ograniczenie dla pieszych i kierowców, określające granice dopuszczalnego ruchu. Sygnalizacja świetlna stanowi ograniczenie, które wskazuje, kiedy dany ruch jest dopuszczalny, a kiedy nie. Przycisk uruchamiający zielone światło dla pieszych, ze względu na taki, a nie inny zamiar projektanta, oferuje przechodniowi wybór takiej możliwości działania, jakim jest wciśnięcie tego przycisku. Z innej strony pewne konwencje kulturowe normują określone zachowania na drodze jako dopuszczalne, tym samym wykluczając inne. Normatywne własności określonych ograniczeń, znaków czy afordancji ściśle wiążą się z funkcjami, jakie pełnią te elementy w szerszych kontekstach. Działania realizowane w takim normatywnie ustrukturyzowanym świecie można określić mianem „praktyk normatywnych" (por. Piekarski, Wachowski, 2018; Wachowski, 2018a).

[67] W proponowanym tutaj ujęciu wszystkie mechanizmy normatywne są jednocześnie mechanizmami przyczynowymi, ale nie wszystkie mechanizmy przyczynowe są jednocześnie normatywne. Normatywne są tylko te mechanizmy, które można określić mianem „konstytutywnych". Dla czynności picia kawy mechanizmem normatywnym będzie taki mechanizm predykcyjny, który odpowiada za wygenerowanie takiej, a nie innej predykcji. Mechanizm ten będzie zatem konstytutywny przyczynowo, bowiem predykcja jest przyczyną konstytutywną takiego, a nie innego działania. Skurcz mięśnia odpowiedzialny za podniesienie kubka kawy ze stołu do ust jest mechanizmem przyczynowym współrealizującym czynność picia kawy, nie jest on jednak mechanizmem konstytutywnym, bowiem kawę można pić na wiele innych sposobów (por. podrozdział *Mechanizmy predykcyjne jako przyczyny a subosobowy poziom wyjaśniania*).

Rozumiem przez to, że praktyki normatywne są takimi praktykami kulturowymi, jakie są realizowane w rozproszonych strukturach normatywnych. Należy dodać, że chodzi tutaj zarówno o działania, które służą realizacji tej normy, jak i jej uniknięciu (por. Wachowski, 2018a, s. 144). Trzeba również pamiętać o tym, że praktyki normatywne mają swój materialny aspekt, czyli fizyczną stronę wykonywanych czynności – określoną lokalizację czasoprzestrzenną, właściwości artefaktów, ich możliwości techniczne (ograniczenia, ulepszenia), dalszy fizyczny kontekst sytuacji (struktura materialna środowiska i innych dostępnych w nim przedmiotów), a także zaangażowanie emocji w użytkowanie rzeczy i udział innych osób. Oznacza to, że relacje normatywne ze światem obejmują również różne rodzaje afordancji, dzięki którym „gotowe" struktury środowiska modyfikują nasze działania bez namysłu lub nawet świadomej kontroli (Piekarski, Wachowski, 2018, s. 410)[68].

Możemy teraz stwierdzić, że normatywność środowiska jest ucieleśniona i osadzona w określonych relacjach normatywnych pomiędzy różnymi własnościami, przedmiotami, zjawiskami i procesami, które są konstytutywne dla działań i zachowań podmiotów. To, co jest charakterystyczne dla głoszonej tutaj hipotezy normatywności rozproszonej, to przekonanie, że nie można normatywności sprowadzić do jakichś określonych przedmiotów czy własności, a że raczej jest odwrotnie: określone przedmioty oraz własności są normatywne na mocy swojej przynależności do szerszej relacyjnej struktury normatywnej. Hipoteza normatywności rozproszonej zakłada zatem pierwotny charakter relacji względem ich elementów. Uwaga ta znajduje swoje uzasadnienie w przeprowadzonych tutaj analizach normatywnych mechanizmów predykcyjnych. Mechanizm jest normatywny na mocy określonych komponentów (tj. ograniczeń środowiskowych, wybranych parametrów modelu generatywnego czy reguły Bayesa), które są konstytutywne dla funkcji przezeń realizowanych. To dopiero jego struktura (relacyjna) czyni dany mechanizm normatywnym, a nie jakiś jeden czy więcej komponentów[69].

5.7. Predykcje skierowane na afordancje

Przeprowadzone tutaj analizy podkreślają ważność powiązania wysokopoziomowych mechanizmów predykcyjnych ze społeczno-materialnymi aspektami środowiska, w którym osadzone są podmioty uzbrojone w model generatywny[70].

[68] Pełna analiza tego aspektu rozproszonej normatywności wykracza poza ramy niniejszych analiz. Jest to z pewnością wątek, który należy przebadać.

[69] Koncepcję normatywności rozproszonej można potraktować jako nowy sposób opisania i być może wyjaśnienia zjawiska normatywności w kontekście społeczno-materialnego środowiska. Jest to ujęcie alternatywne względem rozwijanych ostatnio koncepcji normatywności usytuowanej (*situated normativity*) (por. m.in. Rietveld, 2008). Porównanie obu tych koncepcji wykracza jednak poza ramy badań wyznaczone dla tej pracy.

[70] Twierdzę, że niskopoziomowe mechanizmy predykcyjne można wyjaśnić bez odwoływania się do istnienia afordancji. Wystarczy wskazać określone ograniczenia środowiskowe, np. wzorce, które są przyczyną

Nie jest to oczywiście pomysł zupełnie nowy na gruncie PP, bowiem już wielu badaczy zwróciło uwagę na istotność analiz środowiska dla pełniejszego zrozumienia i wyjaśniania mechanizmów predykcyjnych. Na szczególną uwagę zasługują prace Jelle Bruineberga, Juliana Kiversteina i Erika Rietvelda. Ze względu na doniosłość ich badań, omówię teraz ich stanowisko nazywane „ekologiczno-enaktywnym", jednocześnie wskazując na różnice pomiędzy tym ujęciem a bronioną tutaj wersją PP.

Badacze ci podkreślają, że fristonowska FEP powinna zostać ujęta w kontekście ekologicznym. Zgodnie z FEP rozkład prawdopodobieństwa stanów danego organizmu musi mieć niską entropię, tzn. ma się on znajdować w danym czasie, w jednym określonym stanie. Oznacza to, że wymóg długoterminowego ograniczania entropii jego stanów przekłada się na krótkoterminowy wymóg minimalizowania zaskoczenia, czyli energii swobodnej (por. Friston, 2011). W praktyce oznacza to konieczność ciągłej minimalizacji dywergencji K-L. Zgodnie z założeniami FEP w strukturze organizmu musi odzwierciedlać się struktura jego niszy ekologicznej, tj. środowiska (Friston, 2013a, s. 213):

> podmiot nie ma modelu swojego świata; to raczej on jest modelem. Innymi słowy, forma, struktura i stany naszych wcielonych mózgów nie zawierają modelu sensorycznego – one są tym modelem. […] Ale co to praktycznie znaczy? Oznacza to, że każdy aspekt naszego mózgu można przewidzieć na podstawie naszego środowiska.

Bruinberg z Rietveldem rozumieją to w ten sposób, że FEP umożliwia skuteczne powiązanie dynamiki środowiska z neurodynamiką. Ich zdaniem FEP pozwala sformalizować tzw. tendencję do optymalnej przyczepności (*tendency toward an optimal grip*) w kategoriach dynamicznego sprzężenia dynamiki mózgu z dynamiką całego układu mózg – ciało – środowisko, rozumianego tutaj w terminach krajobrazu afordancji powiązanego z ciałem i mózgiem[71]. Tendencję do optymalnej przyczepności można postrzegać w ramach FEP jako konsekwencję ciągłej minimalizacji energii swobodnej przez percepcję i działania w odniesieniu do organizmu jako całości (Bruineberg, Rietveld, 2014, s. 7). Przez tendencję do optymalnej przyczepności należy rozumieć skłonność indywiduum posiadającego pewne umiejętności do ciągłego doskonalenia swoich reakcji w odniesieniu do takich afordancji, które są istotne dla tego indywiduum w określonej sytuacji. Zatem wewnętrzna dynamika organizmu, który wytwarza model świata w celu skutecznego w nim działania, opiera się na powiązanych z afordancjami wzorcach gotowości do działania (Bruineberg, Rietveld, 2014, s. 6). W praktyce oznacza to,

takiego, a nie innego ustrukturyzowania sygnału zmysłowego. Pełne uzasadnienie tej hipotezy znacznie wykracza poza ramy przyjęte dla tej książki.

[71] W praktyce oznacza to, że należy porzucić myślenie o FEP i PP w terminach inferencjalnych i reprezentacjonistycznych na rzecz podejścia dynamicznego i myślenia o wnioskowaniu bayesowskim w sposób nieobliczeniowy. Omawiani autorzy nie opowiadają się wprost za tymi rozwiązaniami, ale pozostawiają te pytania otwarte, niejako sugerując, że ich podejście może je implikować (por. Bruineberg, Kiverstein, Rietveld, 2016, s. 23).

że to, co jest wywnioskowane w aktywnym wnioskowaniu, to nie są żadne przedmioty ani ich własności, ale raczej wzorce przewidujące, które określają działania konieczne w odniesieniu do odpowiednich afordancji. Takie wzorce nie reprezentują na przykład marchewki, jej długości, kształtu ani tego, co można z nią zrobić, ale raczej stan „atraktora[72]" (organizmu), który jest bezpośrednio powiązany (*coupled*) z dostępnością marchewki i możliwości przezeń oferowanych. Oznacza to, że stan organizmu (modelu generatywnego) jest bezpośrednio powiązany lub innymi słowy – rezonuje z afordancjami oferowanymi przez marchewkę (Dreyfus, 2007, s. 1153).

Pomiędzy organizmem a środowiskiem zachodzi, o czym już wspomnieliśmy, dynamiczne sprzężenie. Opiera się ono na tym, że struktura środowiska, rozumianego tutaj jako krajobraz afordancji, jest zawsze strukturą pola afordancji określonego organizmu. Zależy ona zatem od jego aktualnych umiejętności, możliwości oraz obecnej sytuacji. Struktura pola afordancji zmienia się, gdy zmienia się krajobraz afordancji, np. wtedy gdy zmienia się otoczenie społeczno-materialne lub umiejętności danego organizmu. Jeżeli przykładowo królik zje marchewkę, to wraz ze zmianą w stanie jego organizmu zmieniają się dostępne możliwości działania. Nasycony królik będzie widział norę jako miejsce do spania itd. Zmiany w polu afordancji są zatem zależne zarówno od stanów środowiska, jak i stanów organizmu. Organizm sprawnie funkcjonujący w swoim otoczeniu jest wykwalifikowany intencjonalnie[73]. Reaguje on zatem umiejętnie na określoną afordancję, ale także jest otwarty na zmiany w środowisku i odpowiednie zaangażowanie się w tę afordancję. Taka tendencja jest wynikiem dynamicznego sprzężenia pomiędzy krajobrazem afordancji i obecnym stanem organizmu.

Na podstawie powyższych analiz można powiedzieć, że predykcje generowane przez organizm (rozumiany tutaj jako model swojego środowiska) są ukierunkowane na afordancje. Takie podejście pozwala na alternatywne rozwiązanie problemu ciemnego pokoju i dylematu eksploracji i eksploatacji. W ujęciu Kiversteina, Millera i Rietvelda takie zjawiska jak ciekawość i skłonność do zabawy można ostatecznie sprowadzić do doświadczeń, które mają dodatnią wartość hedoniczną. Nowe doświadczenia, pojawiające się jako efekt naszej ciekawości i chęci do zabawy, są często przeżywane jako przyjemne. W ujęciu Fristona i współpracowników (Friston i in.,

[72] Pojęcie z teorii chaosu oznaczające taki zbiór w przestrzeni fazowej, do którego wraz z upływem czasu zmierzają trajektorie, które rozpoczynają się w różnych obszarach przestrzeni fazowej. Atraktorem może być punkt, zamknięta krzywa bądź fraktal. W ujęciu Bruineberga i Rietvelda atraktor to taki punkt w przestrzeni stanów, do jakiego określony system będzie się kierował w odwołaniu w określonym środowisku. Badacze powołują się na to pojęcie ze względu na jego zastosowanie zaproponowane przez Huberta Dreyfusa. Dreyfus rozumie atraktor jako każdy możliwy minimalny stan energetyczny, który może osiągnąć mózg. W miarę jak aktywacja mózgu zbliża się do odpowiedniego atraktora, mózg w efekcie wybiera znaczący bodziec z tła (Dreyfus, 2007, s. 1152). Pojęcie to również jest wykorzystywane w teorii układów dynamicznych (por. Miller, 2016). Jego wykorzystanie przez omawianych autorów jest tym samym kolejnym krokiem w odchodzeniu od tradycyjnych ujęć obliczeniowych i reprezentacjonistycznych.

[73] Wykwalifikowana intencjonalność (*skilled intentionality*) to rodzaj intencjonalności, którą wykazuje organizm skutecznie działający w swoim środowisku. Taki organizm ma tendencję do optymalnej przyczepności w polu afordancji (Bruineberg, Rietveld, 2014, s. 2).

2017b) uczucie przyjemności powinno być skutkiem rozpoznania takich działań w środowisku (tj. afordancji epistemicznej), jakie powodują minimalizację niepewności, *resp*. maksymalizację wierzytelności modelu. Nie jest jednak zupełnie jasne, dlaczego doświadczenia zmniejszające niepewność wywołują uczucie przyjemności, zaś te które jej nie zmniejszają – uczucie przykrości? Zdaniem omawianych badaczy, dopóki nie odpowie się na to pytanie, dopóty nie wyjaśni się roli, jaką grają wartości w działaniach epistemicznych (Kiverstein, Miller, Rietveld, 2010, s. 2855).

Kiverstein, Miller i Rietveld proponują, aby myśleć o wartościach w kategoriach gotowości do podejmowania działań w reakcji na odpowiednie afordancje. To afordancje mają wartość, a doświadczenia emocjonalne przygotowują podmiot do działania w odniesieniu do odpowiednich afordancji. Wartości zatem zapraszają lub zniechęcają podmiot do podejmowania takich, a nie innych działań w odniesieniu do dostępnych afordancji w środowisku (Kiverstein, Miller, Rietveld, 2019, s. 2858).

Problematyczny w omawianej propozycji jest status eksplanacyjny używanych tutaj takich pojęć jak: tendencja do optymalnej przyczepności czy krajobraz afordancji. Nie jest bowiem jasne, czy są to pojęcia czysto opisowe, które w pewnym sensie oferują utrzymaną w duchu psychologii ekologicznej Gibsona, ale także fenomenologii Merleua-Ponty'ego, redeskrypcję tzw. helmholtzjańskiego PP (por. Bruineberg, 2017); czy może są to jednak pojęcia, które odnoszą się do struktury mechanizmu, *resp*. określonych parametrów w przestrzeni stanów danego układu, co by sugerowało wykorzystanie pojęcia atraktora. Jest to z pewnością alternatywny sposób odczytania PP, zrywający z duchem inferencjalizmu i konstruktywizmu konserwatywnego PP, niemniej w tym przypadku takie kluczowe pojęcia dla bayesowskiego PP jak model generatywny czy właśnie wnioskowanie bayesowskie tracą zupełnie na ostrości i nie jest do końca jasne, jak je należy rozumieć. Na przykład jeżeli modelem generatywnym jest cały organizm (Friston, 2013a, s. 213[74]), a jego zadaniem nie jest reprezentowanie świata, tylko kierowanie interakcjami podmiotu w jego otoczeniu, celem utrzymania równowagi na linii mózg – ciało – środowisko (Bruineberg, Rietveld, 2014, s. 7), to nie jest do końca zrozumiałe, (1) w jaki sposób jest konstytuowana relacja pomiędzy mózgiem a modelem generatywnym, skoro nie są one tym samym; oraz (2) w jaki sposób model może kierować działaniami, skoro nie reprezentuje on w żaden sposób struktury środowiska[75]. Innym problemem jest nieiferencjalne odczytanie wnioskowania Bayesa, które tutaj jest rozumiane w sposób enaktywny jako rodzaj działania (por. Ramstead, Kirchhoff, Friston, 2019), co w zależności od interpretacji może prowadzić do konkluzji, że

[74] „Podmiot nie posiada swojego modelu świata – on jest tym modelem. Innymi słowy: forma, struktura i stany naszych ucieleśnionych mózgów nie zawierają modelu sensorium. One są tym modelem".

[75] Oczywiście w literaturze istnieje wiele antyreprezentacjonistycznych propozycji wyjaśnienia tego problemu, jednak w omawianej koncepcji nie jest jasne, na ile model posługuje się bądź nie posługuje się reprezentacjami, oraz nie jest dobrze wyeksplikowane, w jaki sposób model może kierować działaniami, jeżeli nie reprezentuje określonych rozkładów prawdopodobieństwa.

nie jest ono procesem obliczeniowym, tylko np. procesem kontrolowanym przez dynamikę układu (por. Kelso, 2012)[76].

5.8. Wnioskowanie aktywne jako wnioskowania enaktywne

Związek zachodzący pomiędzy wieloma stanami gotowości do działania na poziomach mikro i makro a afordancjami oparty jest na dynamice samoorganizacji systemu, rozumianej jako enaktywne minimalizowanie energii swobodnej w ramach całościowego systemu podmiot – środowisko. Oznacza to, że ucieleśnione podmioty, tj. modele generatywne, aktywnie wnioskując, zmniejszają swoje niezadowolenie, tzn. kierują nas ku ważnym z perspektywy podmiotu afordancjom i przygotowują na konsekwencje sensoryczne wynikające z kontaktu ze środowiskiem. W takim ujęciu można się zatem zgodzić z ogólną konstatacją Ramsteada, Kirchhoffa i Fristona (2019), że aktywne wnioskowanie jest wnioskowaniem enaktywnym, co oznacza, że jest ono ukierunkowane na wybór i realizację działań przy założeniu sprzężonej i cyklicznej przyczynowości pomiędzy tymi działaniami a percepcją. Oznacza to, że (1) aktywne wnioskowanie jest samoorganizującym się procesem wyboru zasad działania (*policy selection*), (2) model generatywny zawiera w sobie (*entailed*) wewnętrzną dynamikę, (3) a jego stany wewnętrzne kodują model rozpoznania[77] w terminach wystarczających (do działania) statystyk. Przyczynowe znaczenie modelu zasadza się zatem na tym, że gra on określoną rolę w wyborze odpowiednich zasad działania zrelatywizowanych do FEP i fenotypu określonego organizmu. Innymi słowy, modele generatywne są normatywne, ponieważ są modelami, „tego, co powinno być, biorąc pod uwagę rodzaj istoty, którą ja jestem" (Ramstead, Kirchhoff, Friston, 2019, s. 9). Oznacza to, że przyczyniają się one do wyboru takich działań, które utrzymują dany organizm w stanach wyznaczanych przez jego fenotyp[78].

Zaproponowane przez Ramsteada, Kirchhoffa i Fristona rozumienie normatywności wykracza poza dotychczasowe analizy FEP i aktywnego wnioskowania. Tutaj bowiem o normatywnym charakterze modelu nie decydują już tylko normatywne wymogi minimalizacji energii swobodnej oraz konieczność zwiększania wierzytelności modelu, ale także przyczynowa rola, jaką odgrywa model w wyjaśnianiu przyczyn działań. Jest to o tyle istotne, że na tej podstawie można mówić o normatywności

[76] Korbak (2019b) przekonująco pokazuje, broniąc stanowiska, które określa jako „enaktywizm obliczeniowy", że można być zwolennikiem obliczeniowej teorii umysłu i informacji semantycznej, jednocześnie przyjmując przesłanki enaktywizmu i FEP.

[77] Por. Kiefer i Hohwy (2018) na temat tego pojęcia.

[78] Enaktywna interpretacja aktywnego wnioskowania pociąga za sobą wiele istotnych konsekwencji, m.in. odrzucenie znaczenia reprezentacji strukturalnych i konieczności uwzględnienia korespondencji przy wyjaśnianiu znaczenia modelu generatywnego dla skuteczności działań.

modelu w sensie eksplanacyjnym, a nie tylko opisowym, za czym argumentowałem wcześniej, mówiąc o FEP i aktywnym wnioskowaniu. Ujęcie to jednak nie jest satysfakcjonujące, bowiem nie wiadomo dokładnie, co chcemy wyjaśnić, kiedy stwierdzamy, że model jest normatywny. Skoro przez model generatywny rozumie się zasadniczo cały organizm, zaś wszelkie podejmowane przezeń działania są podporządkowane stanom jego fenotypu [ponieważ organizmy oczekują tego, co wyznacza ich fenotyp (Ramstead, Kirchhoff, Friston, 2019, s. 5)], to okazuje się, że normatywne jest to, co jest zgodne z fenotypem organizmu. Z jednej strony mamy tutaj do czynienia z przeformułowaniem FEP, tzn. normatywne są takie zachowania organizmu, jakie spełniają wymóg minimalizacji energii swobodnej, *resp.* unikania śmierci termodynamicznej, czyli takie działania, jakie określa fenotyp, z drugiej zaś trudno jest wyjaśnić możliwość działania nienormatywnego (dysfunkcyjnego). Można oczywiście powiedzieć, i nie jest to żaden błąd, że organizm działa niepoprawnie (tj. nienormatywnie), jeżeli wybiera takie działania, jakie zwiększają niepewność czy zmniejszają wiarygodność modelu, ale niejasnym pozostają dwie kwestie: (1) jak określić kwestię tego, które z zachowań są zgodne z fenotypem i sprowadzalne do adaptacji organizmu w środowisku?[79]; (2) jak wytłumaczyć działania, które przynajmniej czasowo zwiększają niepewność i obniżają wiarygodność modelu? Pierwsza kwestia dotyczy problemu, który pojawił się przy okazji omawiania mechanizmów homeostatycznych: tylko niektóre z naszych niskopoziomowych zachowań można określić mianem „homeostatycznych", pozostałe zaś nie są w żadnym stopniu homeostatyczne i w wielu wypadkach nie da się ich sprowadzić do tych, które są homeostatyczne (por. podrozdział *Normatywność homeostazy*), druga zaś kwestia sprowadza się do omawianego już tutaj problemu ciemnego pokoju i kompromisu eksploatacyjno-eksploracyjnego.

Trzeba zatem stwierdzić, że normatywność sprowadzona wyłącznie do wewnętrznej dynamiki modelu bądź organizmu nie daje się uzasadnić ontycznie, a tylko epistemicznie. Innymi słowy, mówienie o normatywności modelu, przekonań bądź działań sprowadza się wyłącznie do redeskrypcji tych zjawisk w kategoriach optymalne/nieoptymalne, zwiększające/zmniejszające wiarygodność itd. Nie ma ono jednak nietrywialnej wartości eksplanacyjnej. Ujęcia określające z góry, co jest normatywne, a co nie, popełniają analogiczny błąd do tego, który pojawiał się w klasycznych ekonomicznych modelach podejmowania decyzji oraz

[79] Przykładowo: czy latanie samolotem odpowiada fenotypowi człowieka, czy nie? Chodzi o to, że dobrze są nam znane statystyki katastrof lotniczych oraz to, że znacznie bezpieczniej podróżuje się samolotem niż np. samochodem. To znaczy, wiemy, że prawdopodobieństwo poniesienia śmierci w wyniku katastrofy lotniczej jest wielokrotnie mniejsze niż w przypadku jazdy samochodem, ale mimo to odczuwamy większy lęk przed lotem niż przed jazdą samochodem. Jak zatem wytłumaczyć napięcie pomiędzy wysoko poziomymi przekonaniami dotyczącymi bezpieczeństwa latania samolotami a naszym lękiem przed katastrofami lotniczymi? Czy nie jest np. tak, że znajdowanie się na bardzo dużej wysokości zwiększa niepewność, ponieważ taki stan nie odpowiada stanom naszego fenotypu? Można tak powiedzieć, ale przecież nie powiemy z tego powodu, że latanie samolotem jest nieracjonalne, nieoptymalne bądź ostatecznie, że zmniejsza ono wiarygodność modelu itd.

w tradycyjnych modelach Bayesa, mianowicie: sprowadza się wyjaśnienie działań do jakiejś jednej, odgórnie wyznaczonej reguły normatywnej. Takie rozwiązanie jest być może zasadne z punktu widzenia etyki czy prakseologii, niemniej nie jest ono uzasadnione i pożądane w obrębie nauk, których celem jest wyjaśnienie określonych zjawisk i procesów.

Argumentem przemawiającym za antyrealizmem ujęć wykorzystujących ramę aktywnego wnioskowania jest strategia eksplanacyjna, którą Simon McGregor, a za nim niektórzy zwolennicy FEP, określają jako teorię „tak, jakby" („*as if" theory*). Stwierdzenie, że coś zachowuje się „tak, jakby" miało własność *X*, zazwyczaj oznacza, że w rzeczywistości nie ma ono własności *X*, ale że system, któremu przypisujemy własność *X*, musi zachowywać się tak, jakby miał własność *X*. W tym sensie zarówno regulacja temperatury przez termostat, jak i ściganie myszy przez kota jest „tak, jakby" działaniem (McGregor, 2017, s. 72). W odniesieniu do bayesowskiego FEP i PP należałoby powiedzieć, że systemy poznawcze zachowują się tak, jakby wnioskowały o przyczynach pobudzeń zmysłowych na podstawie sygnału zmysłowego lub innymi słowy, działają tak, jakby wnioskowały o najbardziej prawdopodobnym zdarzeniu na podstawie obserwacji zmysłowej. Oznacza to, że każdy system, który unika zaskakujących kontaktów ze światem (tj. zaskakujących stanów sensorycznych), zachowuje się tak, jakby przewidywał i minimalizował uśrednioną energię swobodną w odniesieniu do określonych wektorów czasowych (Ramstead, Kirchhoff, Friston, 2019, s. 4).

Przyjęcie koncepcji „tak, jakby" wyjaśniającej podstawowe mechanizmy działania modelu generatywnego ma określone konsekwencje eksplanacyjne[80]. Po pierwsze wzmacnia bronioną w tej pracy już wcześniej (por. podrozdział *Zastosowanie zasady energii swobodnej w przetwarzaniu predykcyjnym*) hipotezę, zgodnie z którą FEP stanowi jedynie heurystykę badawczą czy ogólną idealizację dla określonych modeli wyjaśniających dane zjawiska. Jeżeli bowiem jest tak, że system poznawczy zachowuje się tak, jakby minimalizował wariacyjną energię swobodną, *resp.* zaskoczenie, to FEP nie może stanowić ogólnej zasady, która ma wyjaśnić organizację systemów biologicznych [od jednokomórkowców po sieci społeczne (por. Friston, 2009, s. 293)]. W tym sensie niektórzy zwolennicy aktywnego wnioskowania (m.in. Casper Hesp, Maxwell Ramstead[81]) przypisują faktyczną wartość eksplanacyjną jedynie pojęciu modelu generatywnego, dodając, że FEP w czysto matematycznej postaci nie ma żadnych zobowiązań ontologicznych i reprezentacyjnych. Oznacza to, że aktywne wnioskowanie nie tyle wyjaśnia określone zjawiska, ile je opisuje, tym samym kształtując odpowiednie „przewodniki epistemiczne" (*epistemological guides*) dla badań empirycznych opisywanych zjawisk[82].

[80] Van Rooij i in. (2018) pokazuje wiele obliczeniowych ograniczeń dla wyjaśnień „tak, jakby".

[81] https://twitter.com/casper_hesp/status/1183647598923702272 (dostęp: 4.11.2019).

[82] Oddzielnym problemem jest to, czy teoretyczne zobowiązania wynikające z FEP dla określonych modeli są charakterystyczne tylko dla FEP? (por. Gładziejewski, w przygotowaniu).

Takie ujęcie sprawy sprowadza się do przyjęcia tzw. słabego preskryptywizmu (epistemicznego), zgodnie z którym modelowanie bayesowskie jest zasadne w odniesieniu do wybranych zjawisk poznawczych z powodów pragmatycznych bądź eksplanacyjnych. Nie przesądza ono jednak, czy dane zjawiska są bayesowskie z natury (por. podrozdział *Preskryptywizm epistemiczny i ontyczny*), tym samym odrzuca się tutaj mocny preskryptywizm (ontyczny), który zakłada, że zjawiska wyjaśnia się przy pomocy twierdzenia Bayesa, ponieważ są one bayesowskie *per se* (por. podrozdział *Normatywność mechanizmów predykcyjnych a preskryptywizm ontyczny*). Tym samym zasadność stosowania FEP opiera się może na różnych względach pragmatycznych czy instrumentalnych, ale także socjologicznych (por. Colombo, Elkin, Hartmann, 2019, s. 24). Ogólnie mówiąc: zastosowanie FEP i aktywnego wnioskowania do wyjaśniania zjawisk poznawczy (szczególnie w PP), a szerzej biologicznych czy społecznych, może prowadzić badaczy do antyrealizmu i różnych form pragmatyzmu bądź instrumentalizmu (Van Es, 2020).

Można oczywiście bronić innej wykładni FEP, zgodnie z którą ma ona określone zobowiązania ontyczne (wydaje się, że takiej interpretacji teorii FEP bronił Friston i inni badacze w ubiegłych latach; por. np. Hohwy, 2013, 2015; Seth, 2015), ale wtedy pojawia się wiele problemów, które omawialiśmy w rozdziale *Zasada energii swobodnej w przetwarzaniu predykcyjnym*, oraz to, co Miłkowski określa mianem „skoku" od hipotezy deskryptywnej dotyczącej tego, co robi mózg, *resp.* model generatywny, do określonych twierdzeń (eksplanacyjnych) na temat takich, a nie innych mechanizmów neuronalnych (Miłkowski, 2018b, s. 12). Innym problemem pozostaje kwestia tego czy FEP, ale również PP, traktowane jako teorie, które nie wymagają integracji z innymi teoriami czy modelami, mają takie moce eksplanacyjne, jakie przypisują im ich zwolennicy (por. podrozdział *Ograniczenia jako warunki integracji modeli*).

Podsumowanie

Celem tego rozdziału była dyskusja nad problemem ciemnego pokoju oraz możliwości jego rozwiązania przez PP i ramę aktywnego wnioskowania. Rozwiązanie tego problemu jest istotne bowiem, jak wykazałem, stanowi w pewnym sensie miarę możliwości eksplanacyjnych określonych podejść. Wywodzące się z FEP aktywne wnioskowanie nie oferuje dobrej odpowiedzi na ten problem z powodu formułowania jednej, ogólnej normatywnej zasady, która ma wyjaśniać wszystkie zachowania i działania podmiotów w kategoriach maksymalizacji wierzytelności modelu, *resp.* minimalizacji zaskoczenia. Argumentowałem za tym, że PP, z powodu przyjętego monizmu eksplanacyjnego, nie daje możliwości nietrywialnego rozwiązania problemu ciemnego pokoju i powiązanego z nim kompromisu eksplo-

racyjno-eksploatacyjnego. Owa trudność jest szczególnie widoczna w przypadku analizy normatywnych mechanizmów predykcyjnych.

Stwierdziłem, że satysfakcjonujące wyjaśnienie mechanizmów normatywnych, kompromisu eksploracji i eksploatacji opierać się musi na odwołaniu do istnienia ograniczeń środowiskowych, bowiem są one tym elementem, za sprawą którego mechanizmy są normatywne. Ograniczenia bowiem nie tylko pełnią funkcję kontekstu dla danego mechanizmu, ale są jego konstytutywnym komponentem. Wykazałem, że warunkiem pełnego wyjaśnienia roli ograniczeń w normatywnych mechanizmach predykcyjnych jest mechanistyczna integracja poszczególnych modeli bądź ram badawczych, bowiem PP zintegrowane z tymi modelami bądź ramami badawczymi umożliwia wyjaśnienie natury normatywnych mechanizmów predykcyjnych.

W ostatniej części tego rozdziału sformułowałem hipotezę normatywności rozproszonej. Zakłada ona, że o normatywnym charakterze określonych mechanizmów, funkcji czy przedmiotów decydują relacje, w które te mechanizmy, funkcje czy przedmiot wchodzą. Innymi słowy, stwierdziłem, że mechanizmy normatywne stanowią element pewnej szerszej struktury normatywnej, która umożliwia danemu mechanizmowi odgrywanie takich, a nie innych ról normatywnych. Na końcu wykazałem słabość eksplanacyjną stanowisk, które określają się jako „enaktywistyczne".

Zakończenie

Celem niniejszej pracy było zarysowanie koncepcji mechanizmów normatywnych oraz próba aplikacji tej koncepcji w odniesieniu do ramy PP. W tym celu przyjąłem pięć hipotez badawczych, które kolejno uzasadniałem:

(H1) Należy odróżnić normatywność oceniającą, tj. tę, która pozwala obserwatorowi w sposób zewnętrzny orzekać o tym, czy dany przedmiot bądź mechanizm jest skuteczny albo nieskuteczny, poprawny bądź niepoprawny; dobry albo zły itd., od normatywności wewnętrznej, czyli tej, którą przypisuje się danemu przedmiotowi czy mechanizmowi z powodów ontycznych (czyli dlatego, że dany przedmiot czy mechanizm jest taki, jaki *jest*). Hipoteza ta nakłada na dany przedmiot bądź mechanizm warunek ontyczny.

(H2) Normatywność jest pojęciem istotnym eksplanacyjnie. Za jego pomocą nie tylko opisujemy, ale także wyjaśniamy określone mechanizmy, wskazując na takie ich (normatywne) własności, jakie są dla nich konstytutywne. Hipoteza ta nakłada na dany przedmiot bądź mechanizm warunek eksplanacyjny.

(H3) Jeżeli istnieją mechanizmy normatywne, to grają one określone role przyczynowe w wyjaśnieniach działań i zachowań. Hipoteza ta nakłada na dany przedmiot bądź mechanizm warunek normatywności (eksplanacyjnej).

(H4) PP oferuje nietrywialne wyjaśnienie określonych zjawisk, wykorzystując modelowanie Bayesa.

(H5) Bayesowskie PP ma charakter normatywny, ponieważ mechanizmy predykcyjne są normatywne (w sensie przedstawionym powyżej), a nie dlatego, że formułuje się tutaj pewne (normatywne) zasady działania lub myślenia.

Wsparciem dla hipotez czwartej i piątej była hipoteza ontyczna (HO), która głosi, że:

(HO) Istnieją takie zjawiska, jakie wyjaśnia się za pomocą bayesowskiego PP, ponieważ są one bayesowskie z natury, a jeżeli są one bayesowskie z natury, to znaczy, że powinno się je wyjaśniać przy użyciu modelowania probabilistycznego.

Zakładając hipotezę ontyczną, stwierdziłem, że hipoteza epistemiczna, którą z wielu powodów można uznać za zasadną, musi ostatecznie zostać odrzucona. Zgodnie z tą hipotezą:

(HE) Istnieją takie zjawiska, jakie wyjaśnia się przy pomocy bayesowskiego PP z powodów pragmatycznych, instrumentalnych, eksplanacyjnych itd.

Struktura książki została podporządkowana realizacji celu, jakim było uzasadnienie wszystkich wymienionych hipotez.

We wprowadzającym rozdziale *Pojęcie przetwarzania predykcyjnego* omówiłem główne pojęcia związane z PP oraz przedstawiłem ogólną charakterystykę tej ramy badawczej. Przedyskutowałem tezę o kantowskim charakterze PP, broniąc poglądu, że w PP nie ma się do czynienia ani z dedukcją transcendentalną, ani z zastosowaniem argumentów transcendentalnych. W PP wykorzystuje się stosowaną w kognitywistyce metodę inżynierii odwrotnej oraz wnioskowania do najlepszego wyjaśnienia (abdukcja). W dalszej części rozdziału zrekonstruowałem zarzut antyrealizmu wobec PP, sformułowany przez Zahaviego. Wykazałem, że charakterystyczne dla konserwatywnego PP stanowisko internalistyczne, zakładające, że granicą systemu poznawczego jest czaszka bądź granica systemu nerwowego, nie implikuje rozstrzygnięć antyrealistycznych, co sugeruje Zahavi. Jest tak, bowiem model generatywny funkcjonuje w ścisłym związku przyczynowym ze światem w taki sposób, że niektóre wzorce statystyczne obecne w sygnale zmysłowym korespondują z realnymi strukturami w świecie. Model reprezentuje je, wykorzystując S-reprezentacje. Następnie wykazałem, korzystając z analiz Drayson, że w PP należy odróżnić trzy znaczenia pośredniości percepcyjnej: psychologiczną, metafizyczną i epistemologiczną. Można twierdzić, że ta rama badawcza zakłada pośredniość tylko w sensie psychologicznym. Nie warunkuje ona jednak w sposób konieczny pośredniości percepcyjnej w sensie metafizycznym bądź epistemologicznym. W dalszej części omówiłem tzw. radykalne wersje PP oraz wskazałem na ich źródła i genezę.

W rozdziale *Przetwarzanie predykcyjne jako bayesowski model eksplanacyjny* argumentowałem za tezą, że bayesowskie PP ma określone moce eksplanacyjne. Wykazałem, że powszechna krytyka modeli bayesowskich jako słabych eksplanacyjnie nie jest zasadna w odniesieniu do PP. Stwierdziłem, że (1) istnieje wiele możliwych algorytmów, które pokazują, w jaki sposób to PP modeluje czynności poznawcze, oraz że (2) istnieją dobre argumenty na rzecz tezy, zgodnie z którą PP ma zastosowanie do mechanistycznego poziomu implementacji. W dalszej części argumentowałem, że bayesowskie PP można uznać za stanowisko bayesizmu oświeconego, tzn. nietrywialnego i eksplanacyjnie satysfakcjonującego, nie zaś za przykład słabego eksplanacyjnie bayesowskiego fundamentalizmu. Następnie

argumentowałem na rzecz tezy, że PP zakłada określoną wersję preskryptywizmu. W tym celu przedyskutowałem tezę Elqayam i Evansa (2011), która dotyczy postulatu porzucenia myślenia o wyjaśnianiu poznania w kategoriach preskryptywnych i nakazuje przejść na pozycje deskryptywne. W odniesieniu do modeli bayesowskich oznacza to, że należy odrzucić przekonanie o ich normatywnym charakterze. Wykazałem, że wspomniani autorzy nie odróżniają normatywności oceniającej, którą wiąże się z powinnością i z określonym systemem ewaluacji, od normatywności wewnętrznej, którą przypisuje się danemu przedmiotowi czy mechanizmowi z powodów ontycznych. W taki sposób została uzasadniona pierwsza hipoteza (H1).

Dyskutując z poglądem Elqayam i Evansa, stwierdziłem, że podział na deskryptywizm bayesowski i preskryptywizm bayesowski jest pozorny i ostatecznie mamy do czynienia z dwoma formami preskryptywizmu, tj. słabą i mocną. Wersja mocna preskryptywizmu zakłada postawę realistyczną w odniesieniu do bayesowskiego PP, zaś wersja słaba implikuje postawę antyrealistyczną. Stwierdziłem, że uzasadnienie realizmu PP jest uzależnione od tego, czy da się wykazać, że mechanizmy poznawcze faktycznie mają charakter bayesowski. Argumentowałem za tym, że słaba wersja preskryptywizmu (preskryptywizm epistemiczny) jest zasadna w odniesieniu do PP. Zakłada ona możliwość wyjaśniania zjawisk przy użyciu twierdzenia Bayesa, nie przesądza jednak, czy zjawiska te są bayesowskie z natury. Następnie wykazałem, że stanowisko realistyczne w odniesieniu do bayesowskiego PP zakłada przyjęcie mocnego (ontycznego) preskryptywizmu. Zgodnie z tym stanowiskiem zjawiska wyjaśnia się za pomocą twierdzenia Bayesa, ponieważ są one bayesowskie. Jeżeli są one bayesowskie z natury, to znaczy, że *powinno* się je wyjaśniać przy użyciu modelowania Bayesa. Tym samym wstępnie uzasadniłem hipotezę drugą (H2) oraz hipotezę epistemiczną (HE), wskazując na konieczność uzasadnienia hipotezy ontycznej (HO).

Rozdział trzeci poświęciłem omówieniu sformułowanej przez Fristona zasady energii swobodnej (FEP). Zgodnie z tą zasadą wszystkie systemy biologiczne (zdefiniowane w terminach koców Markowa) minimalizują energię swobodną swoich stanów wewnętrznych w celu utrzymania homeostazy. Część badaczy twierdzi, że PP jest specjalnym przypadkiem zastosowania tej zasady odniesionym do poznania, zaś mechanizmy predykcyjne są w istocie homeostatycznymi mechanizmami minimalizującymi energię swobodną. Omówienie FEP w kontekście PP było kluczowe, ponieważ zwolennicy tej zasady określają ją jako eksplanacyjną zasadę normatywną. Wydaje się ona zatem kluczowa dla zrozumienia nie tylko normatywnych mechanizmów predykcyjnych, ale normatywnych mechanizmów biologicznych w ogóle. Aby określić możliwości eksplanacyjne FEP, odwołałem się do dyskusji zwolenników tej zasady na temat tzw. problemu ciągłości pomiędzy życiem a umysłem. Skonfrontowałem ze sobą stanowiska kognitywistyczne i nonkognitywistyczne. Krytyczna analiza tej dyskusji oraz dodatkowe sformułowane

przeze mnie argumenty pozwoliły mi na rewizję pretensji eksplanacyjnych FEP oraz odrzucenie przekonania, że jest ona konieczna dla wyjaśnienia natury mechanizmów predykcyjnych. Stwierdziłem, że FEP stanowi istotną heurystykę badawczą dla analiz z zakresu PP. Argumentowałem jednak za tym, że samo PP nie może zostać wyjaśnione w terminach minimalizacji energii swobodnej i koców Markowa.

W rozdziale *Funkcje i mechanizmy normatywne w kontekście przetwarzania predykcyjnego* zastanawiałem się nad normatywnym charakterem FEP. Stwierdziłem, że myślenie o mechanizmach homeostatycznych w kategoriach normatywnych, które zgodnie z ujęciem Fristona są mechanizmami minimalizacji energii swobodnej, nie jest zadowalające. Wykazałem bowiem, że mechanizmy predykcyjne w myśl bronionego tutaj stanowiska są normatywne, jednocześnie nie będąc mechanizmami homeostatycznymi. Argumentowałem za tym, że nietrywialne wyjaśnienie mechanizmów normatywnych powinno wiązać się z wyjaśnieniem natury tzw. funkcji normatywnych. Krytycznie omówiłem obecne w literaturze rozumienia funkcji normatywnych, sprawdzając, czy spełniają one dwa sformułowane przeze mnie wymogi: eksplanacyjny i normatywny. Wykazałem, że żadne z omawianych stanowisk nie spełnia tych wymogów w sposób zadowalający. Następnie odwołałem się do koncepcji Bickharda i stwierdziłem, że (1) specjalnym przypadkiem funkcji normatywnych są funkcje reprezentacyjne i predykcyjne; oraz że (2) teoria Bickharda wymaga zasadniczej modyfikacji z powodu jej antymechanistycznego i termodynamicznego charakteru.

W dalszej części tego rozdziału sformułowałem szkicowy model normatywności dla PP. Podałem argumenty na rzecz przekonania, zgodnie z którym określona funkcja jest normatywna wtedy, gdy jednocześnie (1) pozwala wyjaśnić dysfunkcję danego mechanizmu; (2) przyczynia się do utrzymania stabilności organizmu w ten sposób, że kształtuje i ogranicza możliwe relacje, procesy i zachowania danego systemu; oraz (w odniesieniu do funkcji reprezentowania i funkcji predykcyjnej) (3) gdy można za jej pomocą wyjaśnić możliwość przypisywania wartości logicznych określonym reprezentacjom bądź predykcjom. Mechanizm zaś jest normatywny wtedy, gdy realizuje określone funkcje normatywne i gdy jest możliwe, aby dany system był mechanizmem dla jakiejś aktywności pomimo faktu, że z jakiegoś powodu nie może jej (obecnie) wykonywać. Takie rozumienie normatywności mechanizmów zakłada przyjęcie hipotezy epistemicznej (HE). Stwierdziłem jednak, że nie jest ona satysfakcjonująca poznawczo, i zasugerowałem, że należy przyjąć stanowisko preskryptywizmu ontycznego oraz uzasadnić hipotezę ontyczną (HO). W tym kontekście sformułowałem ontyczną interpretację pojęcia mechanizmu normatywnego: mechanizm albo funkcja jest normatywny wtedy, gdy spełnia on takie, a nie inne funkcje przyczynowe w wyjaśnieniach określonych działań i zachowań. W przypadku normatywnych własności mechanizmów i funkcji predykcyjnych oznacza to, że są one przyczynami określonych działań danego organizmu w środowisku.

Na końcu tego rozdziału argumentowałem za mechanistyczną interpretacją normatywnych funkcji i mechanizmów predykcyjnych. Tym samym mogłem uzasadnić, że (1) PP ma zastosowanie na wszystkich trzech poziomach analizy w rozumieniu Marra; (2) stanowisko preskryptywizmu ontycznego (tj. realizmu) w odniesieniu do PP jest zasadne z powodu ontycznego charakteru wyjaśnień formułowanych w ramach tego podejścia; oraz (3) bayesowski model PP jest nietrywialną perspektywą badawczą ze względu na normatywny charakter bayesowskich mechanizmów predykcyjnych. W ten sposób uzasadniłem hipotezę drugą (H2), trzecią (H3) oraz wstępnie hipotezę piątą (H5). Wykazałem również zasadność przyjęcia hipotezy ontycznej (HO) i odrzucenia hipotezy epistemicznej (HE).

Ostatni rozdział niniejszej rozprawy miał na celu (1) wykazanie mocy eksplanacyjnej zaproponowanego tutaj modelu mechanizmów normatywnych; (2) pokazanie, że oparta na FEP rama aktywnego wnioskowania nie oferuje satysfakcjonującego wyjaśnienia mechanizmów normatywnych oraz działań podmiotów bądź organizmów w środowisku. Z obu tych powodów za oś rozważań w tym rozdziale uznałem problem ciemnego pokoju. Ciemny pokój oznacza stan, w którym mógłby znaleźć się podmiot, gdyby zminimalizował sumę wszystkich potencjalnych błędów predykcyjnych. Wykazałem, że jest to problematyczne, bowiem taki stan wcale nie musi i z zasady nie jest pożądany przez żywe organizmy. Stwierdziłem, że sprawdzianem wartości eksplanacyjnej PP oraz FEP jest możliwość nietrywialnego rozwiązania tego problemu, czyli wykazania, dlaczego podmioty są skłonne wybierać niepewną eksplorację, zamiast nieustannie dążyć do bezpiecznej eksploatacji (kompromis eksploracji i eksploatacji). Analizując kolejne propozycje rozwiązania problemu ciemnego pokoju, pokazałem, że ani powiązana z FEP rama aktywnego wnioskowania, ani PP nie są w stanie podać jego dobrego wyjaśnienia. Następnie wykazałem, że pełne wyjaśnienie mechanizmów normatywnych, a w dalszej kolejności rozwiązanie dylematu eksploracji i eksploatacji zakłada odwołanie się do istnienia ograniczeń obecnych w środowisku. Ograniczenia są bowiem takim elementem otoczenia, jaki czyni dany mechanizm nie tylko przyczynowym, ale także normatywnym. Są one zatem kluczowe dla wyjaśnienia mechanizmów predykcyjnych. Nie pełnią bowiem tylko funkcji kontekstu, w którym mechanizm jest realizowany, ale przede wszystkim (niektóre z nich) są jego konstytutywnymi komponentami. Wykazałem, że pełne wyjaśnienie roli ograniczeń w normatywnych mechanizmach predykcyjnych zakłada integrację poszczególnych modeli określonych zjawisk poznawczych, ponieważ dopiero mechanistyczna integracja PP z innymi modelami pozwoli na nietrywialne i mocne eksplanacyjnie wyjaśnienie natury normatywnych mechanizmów predykcyjnych. Argumentowałem również za tym, że niemożliwość rozwiązania problemu ciemnego pokoju przez PP jest bezpośrednim następstwem monizmu eksplanacyjnego obecnego w tym podejściu.

Broniłem tezy, że bayesowskie PP nie dlatego jest normatywne, że pozwala sformułować określone reguły czy zasady postępowania bądź że zawiera takie zasady, ale dlatego, że normatywne są same mechanizmy predykcyjne, ponieważ warunkują one wybór przez podmioty takich, a nie innych działań. W ten sposób uzasadniłem hipotezę piątą (H5). Wykazując, że mechanizmy normatywne pozwalają na wyjaśnienie kluczowego dla rozwiązania problemu ciemnego pokoju, zjawiska motywacji podmiotów, uzasadniłem hipotezę czwartą (H4). W ten sposób zostały uzasadnione wszystkie hipotezy dotyczące natury i charakteru mechanizmów normatywnych. Tym samym cel tej pracy został zrealizowany.

Przeprowadzone analizy nie wyczerpują z pewnością problemu wyjaśnienia mechanizmów normatywnych. Moja propozycja stanowi raczej wstępny szkic tego, jak należy te biologiczne mechanizmy rozumieć. W tym sensie stanowi ona istotny przyczynek do ogólnej teorii mechanizmów normatywnych. Pozwala bowiem m.in. powiązać systemowe podejście do badania normatywności funkcji z podejściem teleosemantycznym. Należy stwierdzić, że koncepcja zaproponowana w tej pracy jest w pewnym sensie dopasowana do określonego stanowiska teoretycznego, bowiem odwołuje się ona głównie do funkcji poznawczych realizowanych przez dane mechanizmy. Wydaje się jednak, że nietrywialna teoria mechanizmów normatywnych powinna również wyjaśnić innego rodzaju funkcje oraz mechanizmy, jednocześnie spełniając wskazane przeze mnie wymogi[1]. Zamiar ten jednak znacznie wykracza poza ramy przeznaczone dla niniejszych rozważań.

Sądzę, że największym osiągnięciem niniejszych analiz jest (1) zaproponowanie oryginalnej perspektywy myślenia o biologicznych mechanizmach normatywnych w kategoriach normatywności eksplanacyjnej; (2) oraz wykazanie, że zaproponowana tutaj koncepcja pozwala na uzasadnienie tezy, zgodnie z którą istnieje określona przyczynowa rzeczywistość pomiędzy naszymi przekonaniami a działaniami, które podejmujemy na podstawie tych przekonań.

Ważnym uzupełnieniem niniejszych analiz jest spostrzeżenie, że broniona przeze mnie koncepcja mechanizmów predykcyjnych nie zakłada określonych zobowiązań metafizycznych charakterystycznych tylko dla PP. Chodzi o to, że może się przykładowo okazać, że teza mózgu bayesowskiego jest fałszywa i że powinno się wykorzystać jakieś inne alternatywne sposoby modelowania niepewności, takie jak teoria prawdopodobieństwa kwantowego albo rama prawdopodobieństwa precyzyjnego (por. Colombo, Elin, Hartmann, 2018). Oznacza to, że teza, zgodnie z którą istnieją normatywne mechanizmy predykcyjne, nie implikuje wprost tezy, że owe mechanizmy muszą być wyłącznie bayesowskie. W tym sensie teoria mechanizmów normatywnych nie jest teorią bayesowską.

[1] Koncepcje teleosemantyczne przekonująco pokazują, że funkcje normatywne są powiązane z wszelkimi funkcjami, które realizuje dany organizm.

W ostatniej części moich rozważań pokazałem, że mechanizmy normatywne są konstytuowane nie tylko w odniesieniu do odpowiednich funkcji oraz struktury organizmu, ale także w odniesieniu do takich, a nie innych stanów środowiska, które mogą być komponentami takich, a nie innych mechanizmów. Biorąc bowiem pod uwagę ten środowiskowy aspekt mechanizmów normatywnych, wykazałem, że stanowią one element pewnej szerszej struktury normatywnej, którą nazwałem „normatywnością rozproszoną". Hipoteza normatywności rozproszonej zakłada, że o normatywnym charakterze określonych mechanizmów, funkcji czy przedmiotów decydują relacje, w które te mechanizmy, funkcje czy przedmioty wchodzą. Zatem normatywna w sensie pierwotnym jest struktura relacyjna, która konstytuuje normatywność określonych przedmiotów wchodzących w jej skład. Twierdzę, że hipoteza normatywności rozproszonej otwiera przed badaczami zupełnie nowe obszary analizy i pozwala przemyśleć wiele problemów, których tradycyjne ujęcia normatywności nie potrafiły satysfakcjonująco rozwiązać albo których nie rozpoznały. Hipoteza ta, wsparta zaproponowaną tutaj koncepcją mechanizmów normatywnych, wyznacza dalsze kierunki badań i stanowi nietrywialne rozwinięcie analiz tutaj przeprowadzonych.

Bibliografia

Adams, R. A., Brown, H. R., Friston, K. J. (2014). Bayesian inference, predictive coding and delusions. *Avant*, 3(5), 51–88. https://doi.org/10.26913/50302014.0112.0004.

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry*, 4(47), 1–26. https://doi.org/10.3389/fpsyt.2013.00047.

Adams, W. J., Graf, E. W., Ernst, M. O. (2004). Experience can change the 'light-from-above' prior. *Nature Neuroscience*, 7(10), 1057–1058.

Aizawa, K. (2014). The enactivist revolution. *Avant*, 2(5), 19–42. https://doi.org/10.26913/50202014.0109.0002.

Ajdukiewicz, K. (1974). *Logika pragmatyczna*. Warszawa: Wydawnictwo Naukowe PWN.

Allen, M. (2018). The foundation: Mechanism, prediction, and falsification in Bayesian enactivism. *Physics of Life Reviews*, 24, 17–20. https://doi.org/10.1016/j.plrev.2018.01.007.

Allen, M., Friston, K. J. (2018). From cognitivism to autopoiesis: Towards a computational framework for the embodied mind. *Synthese*, 195(6), 2459–2482. https://doi.org/10.1007/s11229-016-1288-5.

Alston, W. P. (1976). Two types of foundationalism. *Journal of Philosophy*, 73, 165–185.

Anderson, J. R. (1990). *The adaptive character of thought*. London: Hillsdale.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, 14, 471–517.

Anderson, M. L. (2014). *After Phrenology: Neural reuse and the interactive brain*. Cambridge: MIT Press.

Anderson, M. L. (2017). Of Bayes and bullets: An embodied, situated, targeting-based account of predictive processing. W: T. Metzinger, W. Wiese (eds.), *Philosophy and Predictive Processing*, 4, 1–14. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573055.

Anderson, M. L., Chemero, T. (2013). The problem with brain GUTs: Conflation of different senses of „prediction" threatens metaphysical disaster. *Behavioral and Brain Sciences*, 36(3), 204–205.

Anscombe, G. E. M. (1957). *Intention*. Cambridge: Harvard University Press.

Anselme, P. (2010). The uncertainty processing theory of motivation. *Behavioural Brain Research*, 208, 291–310.

Ashby, W. R. (1947). Principles of the self-organising dynamic system. *The Journal of General Psychology*, 37, 125–128.

Ashby, W. R. (1963). *Wstęp do cybernetyki* (B. Osuchowska, A. Gosiewski, tłum.). Warszawa: Wydawnictwo Naukowe PWN.

Audi, R. (2009). Ethical naturalism. W: D. Copp (ed.), *The Oxford Handbook of Ethical Theory* (91–121). Oxford: Oxford University Press.

Audi, R. (2013). Natura normatywności oraz projekt naturalizacji pojęć normatywnych. W: B. Brożek, M. Hohol, Ł. Kurek, J. Stelmach (eds.). *W świecie powinności* (37–78). Kraków: Copernicus Center Press.

Austin, J. L. (1993). *Mówienie i poznawanie* (B. Chwedeńczuk, tłum.). Warszawa: Wydawnictwo Naukowe PWN.

Badcock, P. B. (2012). Evolutionary systems theory: A unifying meta-theory of psychological science. *Review of General Psychology*, 16(1), 10–23.

Badcock, P. B., Friston, K. J., Ramstead, M. J. D. (2019). The hierarchically mechanistic mind: A free-energy formulation of the human psyche. *Physics of Life Reviews*, 31, 104–121. https://doi.org/10.1016/j.plrev.2018.10.002.

Badcock, P. B., Friston, K. J., Ramstead, M. J. D., Ploeger, A., Hohwy, J. (2019). The hierarchically mechanistic mind: An evolutionary systems theory of the human brain, cognition and behavior. *Cognitive, Affective & Behavioral Neuroscience*, 19(6), 1319–1351. https://doi.org/10.3758/s13415-019-00721-3.

Ballard, D. H. (1981). Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 2(13), 111–122.

Baltieri, M., Buckley, Ch. L. (2019). The dark room problem in predictive processing and active inference, a legacy of cognitivism? *ALIFE 2019: The 2019 Conference on Artificial Life*, 40–47. https://doi.org/10.1162/isal_a_00137.

Barandiaran, X. E. (2007). Mental life: Conceptual models and synthetic methodologies for a post-cognitivist psychology. W: B. Wallace, A. Ross, J. Davies, T. Anderson (eds.), *The world, the mind and the body: Psychology after cognitivism* (49–90). Exeter: Imprint Academic.

Barandiaran, X. E., Chemero, A. (2009). Animals in the modeling ecosystem. *Adaptive Behavior*, 17(4), 287–292.

Barandiaran, X. E., Di Paolo, E., Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry and spatio-temporality in action. *Journal of Adaptive Behavior*. 17(5), 367–386. https://doi.org/10.1177/1059712309343819.

Barandiaran, X. E., Egbert, M. (2011). Quantifying normative behaviour and precariousness in adaptive agency. W: T. Lenaerts, M. Giacobini i in. (eds.), *Proceedings of the 11th European Conference on Artificial Life* (210–218). Presented at the ECAL: MIT Press.

Barandiaran, X. E., Egbert, M. (2013). Norm-establishing and norm-following in autonomous agency. *Artificial Life Journal*, 20(2014), 5–28.

Barwise, J., Perry, J. (1981). Situations and attitudes. *Journal of Philosophy*, 77, 668–691.

Barwise, J., Perry, J. (1983). *Situations and attitudes*. Cambridge: MIT Press.

Bechtel, W. (2006). *Discovering cell mechanisms: The creation of modern cell biology*. Cambridge: Cambridge University Press.

Bechtel, W. (2008). *Mental mechanisms*: *Philosophical perspectives on cognitive neuroscience*. New York: Routledge.

Bechtel, W. (2011). Mechanism and biological explanation. *Philosophy of Science*, 78, 533–557.

Bechtel, W. (2019). Resituating cognitive mechanisms within heterarchical networks controlling physiology and behavior. *Theory & Psychology*, 29(5), 620–639. https://doi.org/10.1177/0959354319873725.

Bechtel, W., Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36, 421–441.

Bergstrom, C. T., Rosvall, M. (2011). The transmission sense of information. *Biology and Philosophy*, 26, 159–176.

Bermúdez, J. L. (2005). *Philosophy of psychology: A contemporary introduction*. London: Routledge.

Bernard, C. (1865). *Introduction à l'étude de la médecine expérimentale*. Paris: Éditions Garnier-Flammarion.

Bicchieri, C. (2006). *The grammar of society*. New York: Cambridge University Press.

Bicchieri, C., Muldoon, R., Sontuoso, A. (2014). Social Norms. W: E. N. Zalta (ed.), *The Stanford encyclopedia of philosophy* (Spring 2014 Edition). Pobrano z: https://plato.stanford.edu/archives/spr2014/entries/social-norms/ (dostęp: 15.10.2019).

Bickhard, M. H. (2002). The biological emergence of representation. W: T. Brown, L. Smith (eds.), *Emergence and reduction: Proceedings of the 29th annual symposium of the Jean Piaget society* (105–131). Hillsdale: Erlbaum.

Bickhard, M. H. (2003). Process and emergence: Normative function and representation. W: J. Seibt (ed.), *Process theories. Cross disciplinary studies in dynamic* (121–155). Dordrecht: Springer. https://doi.org/10.1007/978-94-007-1044-3_6.

Bickhard, M. H. (2007). Mechanism is not enough. *Pragmatics and Cognition*, 15(3), 573–585.

Bickhard, M. H. (2009). The biological foundations of cognitive science. *New Ideas in Psychology*, 27, 75–84. https://doi.org/10.1016/j.newideapsych.2008.04.001.

Bickhard, M. H. (2014). Model interaktywistyczny (M. Miłkowski, tłum.). *Przegląd Filozoficzno-Literacki*, 2(39), 109–165.

Bickhard, M. H. (2015a). Toward a model of functional brain processes I: Central nervous system functional micro-architecture. *Axiomathes*, 25, 217–238. https://doi.org/10.1007/s10516-015-9275-x.

Bickhard, M. H. (2015b). Toward a model of functional brain processes II: Central nervous system functional macro-architecture. *Axiomathes*, 25, 377–407. https://doi.org/10.1007/s10516-015-9276-9.

Bickhard, M. H. (2016). The anticipatory brain: Two approaches. W: V. C. Müller (ed.), *Fundamental issues of artificial intelligence* (259–281). Berlin: Springer. https://doi.org/10.1007/978-3-319-26485-1_16.

Bickhard, M., Richie, D. M. (1983). *On the nature of representation: A case study of James Gibson's theory of perception*. New York: Praeger.

Bielecka, K. (2018). *Błądzę, więc myślę. Co to jest błędna reprezentacja?* Warszawa: WUW.

Bishop, Ch. M. (2006). *Pattern recognition and machine learning*. New York: Springer-Verlag.

Bock, W. J., von Wahlert, G. (1965). Adaptation and the form-function complex. *Evolution*, 19, 269–299.

Boden, M. A. (2000). Autopoiesis and life. *Cognitive Science Quarterly*, 1, 117–145.

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology*, 76, 198–211. https://doi.org/10.1016/j.jmp.2015.11.003.

Boghossian, P. (1989). The rule-following considerations. *Mind*, 98, 507–549.

Boghossian, P. (2003). The normativity of content. *Philosophical Issues*, 13, 31–45.

Boorse, C. (1976). Wright on functions. *Philosophical Review*, 85, 70–86.

Bowers, J. S., Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, 138(3), 389–414. https://doi.org/10.1037/a0026450.

Braillard, P. A., Malaterre, C. (2015). Explanation in biology. W: P. A. Braillard, Ch. Malaterre (eds.), *History, philosophy and theory of the life sciences* (1–28). Dordrecht: Springer Science + Business Media. https://doi.org/10.1007/978-94-017-9822-8_1.

Brandom, R. B. (1994). *Making it explicit: Reasoning, representing, and discursive commitment*. Cambridge: Harvard University Press.

Brandom, R. B. (2000). *Articulating reasons*. Cambridge: Harvard University Press.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, 47, 139–159.

Broome, J. (2007). Is rationality normative? *Disputatio*, 23, 161–178.

Brożek, B. (2013). Pojęcie normatywności. W: A. Brożek, B. Brożek, J. Stelmach (red.), *Fenomen normatywności* (19–44). Kraków: Copernicus Center Press.

Bruineberg, J. (2017). Active inference and the primacy of the 'I Can'. W: T. Metzinger, W. Wiese (eds.), *Philosophy and Predictive Processing*, 5, 1–18. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573062.

Bruineberg, J., Kiverstein, J., Rietveld, E. (2016). The anticipating brain is not a scientist: The free-energy principle from an ecological-enactive perspective. *Synthese*, 195, 1–28. https://doi.org/10.1007/s11229-016-1239-1.

Bruineberg, J., Rietveld, E. (2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience*, 8(599), 1–14.

Bunke, H, Riese, K. (2008). Graph classification based on dissimilarity space embedding. W: N. da Vitoria Lobo, T. Kasparis, F. Roli, J. T. Kwok, M. Georgiopoulos, G. C. Anagnostopoulos, M. Loog, Marco (eds.), *Structural, syntactic, and statistical pattern recognition. Proceedings. Lecture notes in computer* science, Vol. 5342 (996–1007). Heidelberg: Springer Verlag. https://doi.org/10.1007/978-3-540-89689-0_103.

Burge, T. (2005). Disjunctivism and perceptual psychology. *Philosophical Topics*, 33(1), 1–78.

Burge, T. (2010). *Origins of objectivity*. Oxford: Oxford University Press.

Burge, T. (2011). Disjunctivism again. *Philosophical Explorations*, 14(1), 43–80.

Burr, C. (2017). Embodied decisions and the Predictive Brain. W: T. Metzinger, W. Wiese (eds.), *Philosophy and Predictive Processing*, 7, 1–23. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573086.

Cao, R. (2012). A teleosemantic approach to information in the brain. *Biology and Philosophy*, 27, 49–71. https://doi.org/10.1007/s10539-011-9292-0.

Cassam, Q. (1999). Self-directed transcendental arguments. W: R. Stern (ed.), *Transcendental arguments: Problems and prospects* [b.p.]. Oxford: Oxford University Press.

Chemero, A. (2003). An outline of a theory of affordances. *Ecological Psychology* 15(2), 181–195.

Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge: MIT Press.

Chemero, A. (2012). Modeling self-organization with non well founded set theory. *Ecological Psychology*, 24(1), 46–59.

Chemero, A. (2014). Antyreprezentacjonizm i nastawienie dynamiczne (P. Gładziejewski, tłum.). *Przegląd Filozoficzno-Literacki*, 2(39), 79–107.

Chikofsky, E. J., Cross, J. H. (1990). Reverse engineering and design recovery: A taxonomy. *IEEE Software*, 7(1), 13–17.

Chomsky, N. (1957/2002). *Syntactic structures*. Berlin–New York: Mouton de Gruyter.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: Harvard University Press.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Cambridge: Harvard University Press.

Christensen, W. D. (2012). Natural sources of normativity. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43, 104–112.

Christensen, W. D., Bickhard, M. H. (2002). The process dynamics of normative function. *Monist*, 85(1), 3–28.

Christensen, W. D., Bickhard, M. H. (w przygotowaniu). *Function as design versus function as usefulness*.

Churchland, P. M. (2002). *Mechanizm rozumu, siedlisko duszy. Filozoficzna podróż w głąb mózgu* (Z. Karaś, tłum.). Warszawa: Aletheia.

Churchland, P. S. (2013). *Moralność mózgu. Co neuronauka mówi o moralności?* (M. Hohol, N. Marek, tłum.). Kraków: Copernicus Center Press.

Cisek, P., Kalaska, J. F. (2010). Neural mechanisms for interacting with a world full of action choices. *Annual Review of Neuroscience*, 33, 269–298.

Clark, A. (2013a). Expecting the world: Perception, prediction, and the origins of human knowledge. *Journal of Philosophy*, 110(9), 469–496.

Clark, A. (2013b). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36, 181–204. https://doi.org/10.1017/S0140525X12000477.

Clark, A. (2015a). Radical predictive processing. *The Southern Journal of Philosophy*, 53(S1), 3–27.

Clark, A. (2015b). Predicting peace: The end of the representation wars-A reply to Michael Madary. W: T. Metzinger, J. M. Windt (eds.), *Open MIND*, 7(R), 1–7. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958570979.

Clark, A. (2016). *Surfing uncertainty. Prediction, action and the embodied mind*. Oxford: Oxford University Press.

Clark, A. (2017a). How to knit your own Markov Blanket: Resisting the second law with metamorphic minds. W: T. Metzinger, W. Wiese (eds.), *Philosophy and predictive Processing*, 13, 1–19. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573031.

Clark, A. (2017b). Predictions, precision, and agentive attention. *Consciousness and Cognition*, 56, 115–119.

Clark, A. (2018). A nice surprise? Predictive processing and the active pursuit of novelty. *Phenomenology and the Cognitive Sciences*, 17, 521–534. https://doi.org/10.1007/s11097-017-9525-z.

Clark, A. (2019). Beyond desire? Agency, choice, and the predictive mind. *Australasian Journal of Philosophy*, first online, 1–15. https://doi.org/10.1080/00048402.2019.1602661.

Clark, J. E., Watson, S., Friston, K. J. (2018). What is mood? A computational perspective. *Psychological Medicine*, 48(14), 2277–2284. https://doi.org/10.1017/S0033291718000430.

Collell, G., Fauquet, J. (2015). Brain activity and cognition: A connection from thermodynamics and information theory. *Frontiers in Psychology*, 6(818), 1–11. https://doi.org/10.3389/fpsyg.2015.00818.

Colombo, M., Elkin, E., Hartmann, S. (2018). Being realist about Bayes and the predictive processing theory of mind. *The British Journal for the Philosophy of Science*, axy059, 1–32. https://doi.org/10.1093/bjps/axy059.

Colombo, M., Seriès, P. (2012). Bayes in the brain – on Bayesian modelling in neuroscience. *British Journal for the Philosophy of Science*, 63(3), 697–723.

Colombo, M., Wright, C. (2018). First principles in the life sciences: The free-energy principle, organicism, and mechanism. *Synthese*, first online, 1–26. https://doi.org/10.1007/s11229-018-01932-w.

Conant, R. C., Ashby, R. W. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1, 89–97.

Constant, A., Ramstead, M. J. D., Veissière, S. P. L., Friston, K. (2019). Regimes of expectations: An active inference model of social conformity and human decision making. *Frontiers in Psychology*, 10(679), 1–15. https://doi.org/10.3389/fpsyg.2019.00679.

Copp, D. (2007). Why naturalism? W: D. Copp (ed.), *Morality in a natural world* (33–54). Cambridge: Cambridge University Press.

Copp, D. (2009). Naturalizm moralny i trzy stopnie normatywności (P. Makowski, tłum.). *Etyka*, 42, 51–83. https://doi.org/10.14394/etyka.446.

Corcoran, A. W., Hohwy, J. (2017). *Allostasis, interoception, and the free energy principle: Feeling our way forward*. https://doi.org/10.31234/osf.io/zbqnx.

Corcoran, A. W., Pezzulo, G., Hohwy, J. (2019). From allostatic agents to counterfactual cognisers: Aactive inference, biological regulation, and the origins of cognition. *Preprints*, 1–70, 8.11.2019. https://doi.org/10.20944/preprints201911.0083.v1.

Cosmides, L., Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58, 1–73.

Craik, K. H. (1966). The prospects for an environmental psychology. *IPAR Research Bulletin*, 1, 1–18.

Craver, C. F. (2001). Role functions, mechanisms and hierarchy. *Philosophy of Science*, 68, 31–55.

Craver, C. F. (2007). *Explaining the brain*. Oxford: University Press Oxford.

Craver, C. F. (2009). Mechanisms and natural kinds. *Philosophical Psychology*, 22(5), 575–594.

Craver, C. F. (2012). Functions and mechanisms: A perspectivalist account. W: P. Huneman (ed.), *Functions*. Dordrecht: Springer.

Craver, C. F. (2013). The ontic account of scientific explanation. W: M. I. Kaiser, O. R. Scholz, D. Plenge, A. Hüttemann (eds.), *Explanation in the special sciences: The case of biology and history* (27–52). Dordrecht: Springer Verlag.

Craver, C. F., Kaplan, D. (2018). Are more details better? On the norms of completeness for mechanistic explanations. *British Journal for the Philosophy of Science*, 71(1), 287–319. https://doi.org/10.1093/bjps/axy015.

Craver, C. F., Tabery, J. (2019). Mechanisms in science. W: E. N. Zalta (ed.), *The Stanford encyclopedia of philosophy* (Summer 2019 Edition). Pobrano z: https://plato.stanford.edu/archives/sum2019/entries/science-mechanisms (dostęp: 20.10.2019).

Crisp, R. (1996). Naturalism and non-naturalism in ethics. W: S. Lovibond, S. G. Williams (eds.), *Identity, truth and value* (113–129). Malden: Blackwell Publishers.

Cummins, R. (1975). Functional analysis. *Journal of Philosophy*, 72, 741–764.

Cummins, R. (1996). *Representations, targets, and attitudes*. Cambridge: MIT Press.

Danks, D. (2014). *Unifying the mind: Cognitive representations as graphical models*. Cambridge: The MIT Press.

Darden, L. (2006). *Reasoning in biological discoveries*. Cambridge: Cambridge University Press.

Darden, L., Maull, N. (1977). Interfield theories. *Philosophy of Science*, 44, 43–64.

Davidson, D. (1963). Actions, reasons and causes. *The Journal of Philosophy*, 60(23), 685–686.

Davies, P. S. (2001). *Norms of nature: Naturalism and the nature of functions*. Cambridge: MIT Press.

De Jaegher, H., Di Paolo, E. (2007). Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences*, 6(4), 485–507.

Dennett, D. (1995a). *Darwin's dangerous idea: Evolution and the meanings of life*. New York: Simon & Schuster.

Dennett, D. (1995b). Osobowy i subosobowy poziom wyjaśniania. W: B. Chwedeńczuk (red.), *Filozofia umysłu. Fragmenty filozofii analitycznej* (P. Dziliński, tłum.) (103–109). Warszawa: Wydawnictwo Spacja.

Dennett, D. (2003). Naprawdę przekonani: Strategia intencjonalna i dlaczego ona działa (M. Miłkowski, tłum.). *Przegląd Filozoficzno-Literacki*, 4(6), 87–109.

Dennett, D. (2008). Rzeczywiste wzorce. W: M. Miłkowski, R. Poczobut (red.), *Analityczna metafizyka umysłu* (M. Miłkowski, tłum.) (299–326). Warszawa: Wydawnictwo IFiS PAN.

de Waal, F. (2013). *Małpy i filozofowie* (B. Brożek, M. Furman, tłum.). Kraków: Copernicus Center Press.

de Waal, F. (2014). Natural Normativity: The 'Is' and 'Ought' of Animal Behavior. *Behaviour*, 151, 185–204.

Dewhurst, J. (2017). Folk psychology and the Bayesian brain. W: T. Metzinger, W. Wiese (eds.), *Philosophy And Predictive Processing*, 9, 1–13. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573109.

Devlin, K. (1995). *Logic and information*. Cambridge: Cambridge University Press.

Devitt, M. (2005). Scientific realism. W: F. Jackson, M. Smith (eds.), *The Oxford handbook of contemporary philosophy* (767–791). Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199234769.003.0026.

Di Paolo, E., De Jaegher, H. (2012). The interactive brain hypothesis. *Frontiers in Human Neuroscience*, 6, 163–163. https://doi.org/10.3389/fnhum.2012.00163.

Dotow, D., Nie, L., de Wit, M. M. (2012). Zrozumieć afordancje: przegląd badań nad główną tezą Jamesa J. Gibsona (D. Lubiszewski, N. Strehlau, tłum.). *Avant*, 3(2), 282–295.

Downey, A. (2017). Predictive processing and the representation wars: A victory for the eliminativist (via fictionalism). *Synthese*, 195, 5115–5139.

Drayson, Z. (2014). The personal/subpersonal distinction. *Philosophy Compass*, 9(5), 338–346.

Drayson, Z. (2017). Direct perception and the predictive mind. *Philosophical Studies*, 175, 3145–3164. https://doi.org/10.1007/s11098-017-0999-x.

Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge: MIT Press.

Dretske, F. (2004). *Naturalizowanie umysłu* (B. Świątczak, tłum.). Warszawa: Wydawnictwo IFIS PAN.

Dretske, F. (2014). Błędna reprezentacja (K. Bielecka, tłum.). *Przegląd Filozoficzno-Literacki*, 2(39), 39–58.

Dreyfus, H. L. (2007). Why heideggerian AI failed and how fixing it would require making it more heideggerian. *Philosophical Psychology*, 20, 247–268. https://doi.org/10.1080/09515080701239510.

Duchliński, P. (2015). Normatywność – jej źródła i sposoby badania. Analiza metateoretyczna. W: P. Duchliński, A. Kobyliński, R. Moń, E. Podrez (red.), *O normatywności w etyce* (16–62). Kraków: Wydawnictwo WAM.

Easwaran, K. (2011). Bayesianism II: Applications and criticisms. *Philosophy Compass*, 6(5), 321–332. https://doi.org/10.1111/j.1747-9991.2011.00398.x.

Eckstein, M., Collins, E. (2019). *Computational evidence for hierarchically-structured reinforcement learning in humans* (preprint). bioRxiv, https://doi.org/10.1101/731752.

Egan, F. (2014). How to think about mental content. *Philosophical Studies*, 170, 115–135. https://doi.org/10.1007/s11098-013-0172-0.

Elqayam, S., Evans, J. S. (2011). Subtracting „ought" from „is": Descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences*, 34(5), 233–248. https://doi.org/10.1017/S0140525X1100001X.

Enoch, D. (2006). Agency, Schmagency: Why normativity won't come from what is constitutive of action. *Philosophical Review*, 115, 169–198.

Evans, M., Grimm, V., Johst, K., Knuuttila, T., de Langhe, R., Lessells, C., Merz, M., O'Malley, M., Orzack, S., Weisberg, M., Wilkinson, D., Wolkenhauer, O., Benton, T. (2013). Do simple models lead to generality in ecology? *Trends Ecology Evolution*, 28(10), 578–583. https://doi.org/10.1016/j.tree.2013.05.022.

Ey, H. (1956). *La conscience*. Paris: PUF.

Feldman, H., Friston, K. J. (2010). Attention, uncertainty and free-energy. *Frontiers in Human Neuroscience*, 4(215), 1–23. https://doi.org/10.3389/fnhum.2010.00215.

Felleman, D., Van Essen, D. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1), 1–47.

Festinger, L. (1957). *A theory of cognitive dissonance*. California: Stanford University Press.

Fink, S. B., Zednik, C. (2017). Meeting in the dark room: Bayesian rational analysis and hierarchical predictive coding. W: T. Metzinger, W. Wiese (eds.), *Philosophy and Predictive Processing*, 8, 1–13. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573154.

Firth, R. (1964). Coherehce, certainty and epistemic priority. *Journal of Philosophy*, 61, 545–557.

Fishburn, P. C. (1988). Normative theories of decision making under risk and under uncertainty. W: J. Kacprzyk, M. Roubens (eds.), *Non-conventional preference relations in decision making. Lecture notes in economics and mathematical systems*, Vol. 301 (1–21). Berlin–Heidelberg: Springer.

Fiske, S. T., Taylor, S. E. (2008). *Social cognition: From brains to culture*. Boston: McGraw-Hill Higher Education.

Fitelson, B. (2006). The paradox of confirmation. *Philosophy Compass*, 1(1), 95–113. https://doi.org/10.1111/j.1747-9991.2006.00011.x.

FitzGerald, T., Dolan, R., Friston, K. J. (2015). Dopamine, reward learning, and active inference. *Frontiers in Computational Neuroscience*, 9(136), 1–16. https://doi.org/10.3389/fncom.2015.00136.

FitzPatrick, W. J. (2011). Ethical non-naturalism and normative properties. W: M. Brady (ed.), *New waves in metaethics. New waves in philosophy* (7–35). London: Palgrave Macmillan.

Friston, K. J. (2003). Learning and inference in the brain. *Neural Networks*, 16(9), 1325–1352. https://doi.org/10.1016/j.neunet.2003.06.005.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 360, 815–836. https://doi.org/10.1098/rstb.2005.1622.

Friston, K. J. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, 13(7), 293–301.

Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11, 127–138.

Friston, K. J. (2011). Embodied inference: Or „I think therefore I am, if I am what I think". W: W. Tschacher, C. Bergomi (eds.), *The implications of embodiment (cognition and communication*) (1–37). Exeter: Imprint Academic.

Friston, K. J. (2012). A free energy principle for biological systems. *Entropy*, 14, 2100–2121. https://doi.org/10.3390/e14112100.

Friston, K. J. (2013a). Active inference and free energy. *Behavioral and Brain Sciences*, 36(3), 212–213. https://doi.org/10.1017/S0140525X12002142.

Friston, K. J. (2013b). Consciousness and hierarchical inference. *Neuropsychoanalysis*, 15(1), 38–42.

Friston, K. J. (2013c). Life as we know it. *Journal of The Royal Society Interface*, 10, [b.p.]. https://doi.org/10.1098/rsif.2013.0475.

Friston, K. J., Adams, R. A., Montague, R. (2012). What is value – accumulated reward or evidence? *Frontiers in Neurorobotics*, 6(11), 1–25. https://doi.org/10.3389/fnbot.2012.00011.

Friston, K. J., Adams, R. A., Perrinet, L., Breakspear, M. (2012a). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology*, 3, 1–20.

Friston, K. J., Daunizeau, J., Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLoS ONE*, 4(2009), 1–13. https://doi.org/10.1371/journal.pone.0006421.

Friston, K. J., Daunizeau, J., Kilner, J., Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. *Biological Cybernetics*, 102, 227–260. https://doi.org/10.1007/s00422-010-0364-z.

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P. O., Doherty, J., Pezzulo, G. (2016). Active inference and learning. *Neuroscience & Biobehavioral Reviews*, 68, 862–879. https://doi.org/10.1016/j.neubiorev.2016.06.022.

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G. (2017a). Active inference: A process theory. *Neural Computation*, 29(1), 1–49.

Friston, K. J., Fortier, M., Friedman, D. A. (2018). Of woodlice and men: A Bayesian account of cognition, life and consciousness. An interview with Karl Friston. *ALIUS Bulletin*, 2, 17–43.

Friston, K. J., Harrison, L., Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, 19, 1273–1302. https://doi.org/10.1016/S1053-8119(03)00202-7pmid:12948688.

Friston, K. J., Kiebel, S. J. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B*, 364, 1211–1221.

Friston, K. J., Kilner, J., Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1–3), 70–87.

Friston, K. J., Levin, M., Sengupta, B., Pezzulo, G. (2015a). Knowing one's place: A free-energy approach to pattern regulation. *Journal of the Royal Society Interface*, 12(105), 1–12.

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., Ondobaka, S. (2017b). Active inference, curiosity and insight. *Neural Computation*, 29, 2633–2683. https://doi.org/10.1162/neco_a_00999.

Friston, K. J., Mattout, J., Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, 104, 137–160.

Friston, K. J., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., Pezzulo, G. (2015b). Active inference and epistemic value. *Cognitive Neuroscience*, 6(4), 187–214.

Friston, K. J., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., Dolan, R. J. (2014). The anatomy of choice: Dopamine and decision-making. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1655), 1–12.

Friston, K. J., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., Dolan, R. J., Moran, R., Stephan, K. E., Bestmann, S. (2012b). Dopamine, affordance and active inference. *PLOS Computational Biology*, 8(1), 1–20.

Friston, K. J., Stephan, K. E. (2007). Free energy and the brain. *Synthese*, 159, 417–458.

Friston, K. J., Thornton, C., Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, 3(130), [b.p.]. https://doi.org/10.3389/fpsyg.2012.00130.

Froese, T. (2014). Review of radicalizing enactivism: Basic minds without content by D. D. Hutto, E. Myin. *The Journal of Mind and Behavior*, 35, 1–2.

Gallagher, S. (2017). *Enactivist interventions: Rethinking the mind*. Oxford: Oxford University Press.

Gallagher, S., Allen, M. (2016). Active inference, enactivism and the hermeneutics of social cognition. *Synthese*, 195, 2627–2648. https://doi.org/10.1007/s11229-016-1269-8.

Garson, J. (2013). Functional sense of mechanism. *Philosophy of Science*, 80, 317–333. https://doi.org/10.1086/671173.

Garson, J. (2016). *A critical overview of biological functions*. Berlin: Springer.

Gawande, A. (2010). *The checklist manifesto: How to get things right*. NewYork: Henry Holt and Company.

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, 59, 167–192. https://doi.org/10.1146/annurev.psych.58.110405.085632.

Geisler, W. S., Diehl, R. L. (2003). A Bayesian approach to the evolution of perceptual and cognitive systems. *Cognitive Science*, 27, 379–402.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston: Houghton Mifflin.

Gibson, J. J. (1967). New reasons for realism. *Synthese*, 17(1), 162–172.

Gibson, J. J. (1979). *The ecological approach to visual perception*. New York: Psychology Press.

Giere, R. (2001). Critical hypothetical evolutionary naturalism. W: C. Heyes, D. L. Hull (eds.), *Selection theory and social construction* (53–70). Albany: State University of New York Press.

Gigerenzer, G., Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1(1), 107–143. https://doi.org/10.1111/j.1756-8765.2008.01006.x.

Glanville, R. (2003). Second order cybernetics. W: F. Parra-Luna (ed.), *Systems science and cybernetics*. W: *Encyclopedia of life support systems (EOLSS), Developed under the Auspices of the UNESCO* (1–28). Oxford: Eolss Publishers.

Glanville, R. (2004). The purpose of second-order cybernetics. *Kybernetes*, 33(9/10), 1379–1386. https://doi.org/10.1108/03684920410556016.

Glennan, S., Illari, P. (eds.) (2018). *The Rutledge handbook of mechanisms and mechanical philosophy*. London and New York: Routledge.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge: MIT Press.

Gładziejewski, P. (2015). *Wyjaśnianie za pomocą reprezentacji mentalnych*. Toruń: Wydawnictwo UMK.

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, 193, 559–582. https://doi.org/10.1007/s11229-015-0762-9.

Gładziejewski, P. (2017a). The evidence of the senses – a predictive processing-based take on the Sellarsian dilemma. W: T. Metzinger, W. Wiese (eds.), *Philosophy and Predictive Processing*, 15, 1–15. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573161.

Gładziejewski, P. (2017b). Just how conservative is conservative predictive processing? *Hybris*, 38, 98–122.

Gładziejewski, P. (2019). Mechanistic unity and the predictive mind. *Theory & Psychology*, 29(5), 657–675. https://doi.org/10.1177/0959354319866258.

Gładziejewski, P. (w przygotowaniu). *Free energy principle and predictive processing: Disentantling their theoretical commitments*.

Gładziejewski, P., Miłkowski, M. (2017). Structural representations: Causally relevant and different from detectors. *Biology & Philosophy*, 32, 337–355. https://doi.org/10.1007/s10539-017-9562-6.

Godfrey-Smith, P. (1993). Functions: Consensus without unity. *Pacific Philosophical Quarterly*, 74, 196–208.

Godfrey-Smith, P. (1996). *Complexity and the function of mind in nature*. Cambridge: University Press.

Godfrey-Smith, P. (2007). Information in biology. W: D. Hull, M. Ruse (eds.), *The Cambridge companion to the philosophy of biology* (103–119). Cambridge: Cambridge University Press.

Godfrey-Smith, P. (2011). Review of signals, by Brian Skyrms. *Mind*, 120, 1288–1297.

Godfrey-Smith, P. (2018). *Inne umysły. Ośmiornice i prapoczątki świadomości* (M. Adamiec-Siemiątkowski, tłum.). Kraków: Copernicus Center Press.

Godfrey-Smith, P., Sterelny, K. (2016). Biological information. W: E. N. Zalta (ed.), *The Stanford encyclopedia of philosophy* (Summer 2016 Edition). Pobrano z: https://plato.stanford.edu/archives/sum2016/entries/information-biological/ (dostęp: 15.07.2019).

Goldman, A. (1977). *A theory of human action*. Princeton: Princeton University Press.

Goldman, A. (1986). What is justified belief? W: P. Moser (ed.), *Empirical knowledge* (171–192). Lanham: Rowman & Littlefield Publishers.

Gordon, N., Tsuchiya, N., Koenig-Robert, R., Hohwy, J. (2018). *Expectation and attention increase the integration of top-down and bottom-up signals in perception through different pathways* (preprint). bioRxiv, 446948.

Gottlieb, J., Oudeyer, P. Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, 19, 758–770.

Graves, A., Clancy, K. (2019). Pobrano z: https://deepmind.com/blog/article/unsupervised-learning (dostęp: 5.09.2019).

Green, D. M., Swets, J. A. (1988). *Signal detection and psychophysics* (reprint ed.). Los Altos: Peninsula Publishing.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 290, 181–197.

Griffith, T. W., Nersessian, N. J., Goel, A. K. (1996). The role of generic models in conceptual change. W: G. W. Cottrel (ed.), *Proceedings of the 18th Annual Conference of the Cognitive Science Society* (312–317). California: University of California.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., Tenenbaum, J. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8), 357–364.

Griffiths, T. L., Lieder, F., Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, 7, 217–229.

Griffiths, T. L., Kemp, C., Tenenbaum, J. B. (2008). Bayesian models of cognition. W: R. Sun (ed.), *The Cambridge handbook of computational cognitive modeling* (1–49). Cambridge: Cambridge University Press.

Grobler, A. (2019). *Epistemologia. Sandwiczowa teoria wiedzy*. Kraków: Universitas.

Guo, X. i in. (2016). *Deep learning for reward design to improve Monte Carlo tree search in ATARI games* (preprint). arXiv, arXiv:1604.07095.

Hahn, U. (2014). The Bayesian boom: Good thing or bad? *Frontiers in Psychology*, 5(765), 1–12. https://doi.org/10.3389/fpsyg.2014.00765.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814.

Harkness, D. L. (2015). From explanatory ambition to explanatory power – a commentary on Jakob Hohwy. W: T. Metzinger, J. M. Windt (eds.), *Open MIND*, 19(C) [b.p.]. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958570153.

Harkness, D. L., Keshava, A. (2017). Moving from the what to the how and where – Bayesian models and predictive processing. W: T. Metzinger, W. Wiese (eds.), *Philosophy and Predictive Processing*, 16, 1–10. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573178.

Harman, G. (1973). *Thought*. Princeton: Princeton University Press.

Harrison, C. W. (1952). Experiments with linear prediction in television. *Bell System Technical Journal*, 31(4), 764–783.

Hartmann, N. (1926). *Ethik*. Berlin: De Gruyter.

Haugeland, J. (1993). Pattern and being. W: B. Dahlbom (ed.), *Dennett and his critics* [b.p.]. Cambridge: Blackwell.

Heinrich, D. (1969). The proof-structure of Kant's transcendental deduction. *Review of Metaphysics*, 22(4), 640–659.

Heller, M., Lubański, M., Ślaga, S. (1997). *Zagadnienia filozoficzne współczesnej nauki*. Warszawa: Wydawnictwo ATK.

Helmholtz, H. v. (1867). *Handbuch der physiologischen optik*. Leipzig: Leopold Voss.

Helmholtz, H. v. (1995). *Science and culture: Popular and philosophical lectures*. Chicago: Chicago University Press.

Hempel, C. (1965). *Aspects of scientific explanation*. New York: Free Press.

Henderson, D. (2017). Norms. W: H. Kingcaid (ed.), *The Oxford handbook of philosophy of social science* (409–435). Oxford: Oxford University Press.

Heras-Escribano, M., de Pinedo, M. (2016). Are affordances normative? *Phenomenology and the Cognitive Sciences*, 15(4), 565–589.

Hesp, C., Ramstead, M. J. D., Constant, A., Badcock, P., Kirchhoff, M., Friston, K. J. (2019). A multi-scale view of the emergent complexity of life: A free-energy proposal. W: G. Georgiev, J. Smart, C. Flores Martinez, M. Price (eds.), *Evolution, development and complexity: Multiscale models in complex adaptive systems* [b.p.]. Springer Proceedings in Complexity. https://doi.org/10.1007/978-3-030-00075-2_7.

Hildreth, E. C., Ullman, S. (1989). The computational study of vision. W: M. Posner (ed.), *Foundations of cognitive science* (581–630). Cambridge: MIT Press.

Hinton, G. E., Sejnowski, T. J. (1983). Optimal perceptual inference. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, [b.p.]. Washington.

Hipólito, I., Ramstead, M. J. D., Convertino, L., Bhat, A., Friston, K. J., Parr, T. (2020). *Markov blankets in the brain* (preprint). arXiv:2006.02741 [q-bio.NC].

Hohwy, J. (2007). Functional integration and the mind. *Synthese*, 159(3), 315–328.

Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.

Hohwy, J. (2014). The self-evidencing brain. *Noûs*, 50(2), 259–285.

Hohwy, J. (2015a). The neural organ explains the mind. W: T. Metzinger, J. M. Windt (eds.), *Open MIND*, 19(T), 1–22. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958570016.

Hohwy, J. (2015b). The diversity of Bayesian explanation – a reply to Dominic L. Harkness. W: T. Metzinger, J. M. Windt (eds.), *Open MIND*, 19(R), 1–6. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958570870.

Hohwy, J. (2017). How to entrain your evil demon. W: T. Metzinger, W. Wiese (eds.), *Philosophy and Predictive Processing*, 2, 1–15. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573048.

Hohwy, J. (2018). The predictive processing hypothesis. W: A. Newen, L. S. de Bruin, S. Gallagher (eds.), *The Oxford handbook of 4e cognition* (129–145). Oxford: Oxford University Press.

Hohwy, J. (2020a). New directions in predictive processing. *Mind & Language*, 2(35), 209–223. https://doi.org/10.1111/mila.12281.

Hohwy, J. (2020b). Self-supervision, normativity and the free energy principle. *Synthese*, first online, 1–25. https://doi.org/10.1007/s11229-020-02622-2.

Hohwy, J., Roepstorff, A., Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, 108(3), 687–701.

Hookway, C. (1999). Modest sceptical arguments and sceptical doubts: A reply to Stroud. W: R. Stern, *Transcendental arguments: Problems and prospects* (173–187). Oxford: Oxford University Press.

Hopfield, J. J. (1994). Physics, computation, and why biology looks so different. *Journal of Theoretical Biology*, 171(1), 53–60.

Höffe, O. (2003). *Immanuel Kant* (A. Kaniowski, tłum.). Warszawa: Wydawnictwo Naukowe PWN.

Hume, D. (2007). *Traktat o naturze ludzkiej* (C. Znamierowski, tłum.). Warszawa: Aletheia.

Hurley, S. (1998). *Consciousness in action*. Cambridge: Harvard University Press.

Hurley, S. (2001). Perception and action: Alternative views. *Synthese*, 1(129), 3–40. https://doi.org/10.1023/A:1012643006930.

Husserl, E. (1974). *Idee czystej fenomenologii i fenomenologicznej filozofii. Księga druga* (D. Gierulanka, tłum.). Warszawa: Wydawnictwo Naukowe PWN.

Husserl, E. (1982). *Medytacje kartezjańskie z dodaniem uwag krytycznych Romana Ingardena* (A. Wajs, tłum.). Warszawa: Wydawnictwo Naukowe PWN.

Hutchins, E. (1995). *Cognition in the wild*. Cambridge: MIT Press.

Hutchins, E. (2014). The cultural ecosystem of human cognition. *Philosophical Psychology*, 27, 34–49. https://doi.org/10.1080/09515089.2013.830548.

Hutto, D. (2005). Knowing what? Radical versus conservative enactivism. *Phenomenology and the Cognitive Sciences*, 4, 389–405.

Hutto, D. (2006). Against passive intellectualism: Reply to Crane. W: R. Menary (ed.), *Radical enactivism* (121–149). London: John Benjamins Publishing Company.

Hutto, D., Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. Cambridge: MIT Press.

Hutto, D., Myin, E. (2017). *Evolving enactivism: Basic minds meet content*. Cambridge: MIT Press.

Ieropoulos, I., Melhuish, C., Greenman, J. (2003). Artificial metabolism: Towards true energetic autonomy in artificial life. W: W. Banzhaf, I. Ziegler, T. Christaller, P. Dittrich, J. T. Kim (eds.), *Advances in artificial life. ECAL 2003. Lecture notes in computer science*, Vol. 2801 (792–799). Berlin, Heidelberg: Springer.

Illari, P. M. (2013). Mechanistic explanation: Integrating the ontic and epistemic. *Erkenntnis*, 78 (2), 237–255.

Illari, P. M., Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science*, 2, 119–135. https://doi.org/10.1007/s13194-011-0038-2.

Ingarden, R. (1989). *Wykłady z etyki*. Warszawa: Wydawnictwo Naukowe PWN.

Isaac, A. (2019). The semantics latent in Shannon information. *The British Journal for the Philosophy of Science*, 1(70), 103–125. https://doi.org/10.1093/bjps/axx029.

Jasrasaria, D., Pyzer-Knapp, E. (2018). *Dynamic control of explore/exploit trade-off in Bayesian optimization* (preprint). arXiv:1807.01279v1.

Jensen, K., Call, J., Tomasello, M. (2007). Chimpanzees are rational maximizers in an ultimatum game. *Science*, 318, 107–109. https://doi.org/10.1126/science.1145850.

Jensen, F. V. (2006). *An introduction to Bayesian networks*. London: UCL Press.

Jones, M., Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and brain sciences*, 34, 169–231. https://doi.org/10.1017/S0140525X10003134.

Juechems, K., Summerfield, Ch. (2019). Where does value come from? *Trends in Cognitive Sciences*, 23(10), 836–850.

Juś, M. (2014). *Spór o redukcjonizm w medycynie. Studium filozoficzne i metodologiczne*. Toruń: Wydawnictwo UMK.

Kaaronen, R. O. (2018). A theory of predictive dissonance: Predictive processing presents a new take on cognitive dissonance. *Frontiers in Psychology*, 9, 2218. https://doi.org/10.3389/fpsyg.2018.02218.

Kahneman, D. (2012). *Pułapki myślenia. O myśleniu szybkim i wolnym* (P. Szymczak, tłum.). Poznań: Media Rodzina.

Kahneman, D., Tversky, A. (1974). Judgment under uncertainty: Heuristics and Biases. *Science*, 4157(185), 1124–1131. https://doi.org/10.1126/science.185.4157.1124.

Kahneman, D., Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 2(47), 263–291.

Kahneman, D., Tversky, A. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 4(5), 297–323.

Kalinowski, J. (1960). *Teoria poznania praktycznego*. Lublin: WN KUL.

Kant, I. (2001). *Krytyka czystego rozumu* (R. Ingarden, tłum.). Kęty: Antyk.

Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese*, 183, 339–373. https://doi.org/10.1007/s11229-011-9970-0.

Kaplan, D. M. (2012). How to demarcate the boundaries of cognition. *Biology and Philosophy*, 27(4), 545–570. https://doi.org/10.1016/j.shpsa.2018.05.013.

Kaplan, D. M., Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science*, 78, 601–627.

Kauffman, S. (1993). *The origins of order: Self-organization and selection in evolution*. Oxford: Oxford University Press.

Kawalec, P. (2000). Bayesianizm. W: P. Gondek, A. Maryniarczyk, M. Nawracała-Urban, Z. Pańpuch, P. Skrzydlewski, M. St. Zięba (red.), *Powszechna encyklopedia filozofii* (507–509). Lublin: PTTA.

Kawalec, P. (2003). Zagadnienia metodologiczne w bayesowskiej teorii konfirmacji. *Roczniki Filozoficzne*, 1(51), 113–142.

Kawalec, P. (2006). *Przyczyna i wyjaśnianie. Studium z filozofii i metodologii nauk*. Lublin: Wydawnictwo KUL.

Kawalec, P. (2012). Bayesianizm w polskiej tradycji probabilizmu – studium stanowiska Kazimierza Ajdukiewicza. *Ruch Filozoficzny*, 1(69), 111–122.

Keijzer, F., van Duijn, M., Lyon, P. (2013). What nervous systems do: Early evolution, input-output, and the skin brain thesis. *Adaptive Behavior*, 21(2), 67–85. https://doi.org/10.1177/1059712312465330.

Kelsen, H. (1945). *General theory of law and state*. Cambridge: Harvard University Press.

Kelso, J. S. (2012). Multistability and metastability: Understanding dynamic coordination in the brain. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 367, 906–918. https://doi.org/10.1098/rstb.2011.0351.

Kiebel, S., Daunizeau, J., Friston, K. J. (2009). Perception and hierarchical dynamics. *Frontiers in Neuroinformatics*, 3(20), 1–9. https://doi.org/10.3389/neuro.11.020.2009.

Kiebel, S., Friston, K. J. (2011). Free energy and dendritic self-organization. *Frontiers in Systems Neuroscience*, 5(80), 1–13. https://doi.org/10.3389/fnsys.2011.00080.

Kiefer, A. (2017). Literal perceptual inference. W: T. Metzinger, W. Wiese (eds.), *Philosophy and Predictive Processing*, 17, 1–19. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573185.

Kiefer, A., Hohwy, J. (2017). Content and misrepresentation in hierarchical generative models. *Synthese*, 195, 2387–2415.

Kim, J. (1988). What is 'naturalized epistemology'? *Philosophical Perspectives*, 2, 381–405.

Kirchhoff, M. (2018). Autopoiesis, free energy, and the life–mind continuity thesis. *Synthese*, 195, 2519–2540.

Kirchhoff, M., Froese, T. (2017). Where there is life there is mind: In support of a strong life-mind continuity thesis. *Entropy*, 19(4), 1–18. https://doi.org/10.3390/e19040169.

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., Kiverstein, J. (2018). The Markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, 15, 1–11. https://doi.org/10.1098/rsif.2017.0792.

Kirchhoff, M., Robertson, I. (2018). Enactivism and predictive processing: A non-representational view. *Philosophical Explorations*, 2(21), [b.p.]. https://doi.org/10.1080/13869795.2018.1477983.

Kitcher, P. (1993). Function and design. *Midwest Studies in Philosophy*, 1(18), 379–397.

Kiverstein, J., Miller, M., Rietveld, E. (2019). The feeling of grip: Novelty, error dynamics, and the predictive brain. *Synthese*, 196, 2847–2869. https://doi.org/10.1007/s11229-017-1583-9.

Klawiter, A. (2012). Co ze mną zrobisz, kiedy mnie zobaczysz? Percepcja jako wyszukiwanie ofert (affordances) w otoczeniu. *Avant*, 2(3), 261–266.

Klein, C. (2018). What do predictive coders want? *Synthese*, 195, 2541–2557. https://doi.org/10.1007/s11229-016-1250-6.

Knill, D., Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27, 712–719.

Kolchinsky, A., Wolpert, D. H. (2018). Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus* (preprint). https://doi.org/10.1098/rsfs.2018.0041. arXiv:1806.08053v3.

Korbak, T. (2015). Scaffolded minds and the evolution of content in signaling pathways. *Studies in Logic, Grammar and Rhetoric*, 41(1), 89–103. https://doi.org/10.1515/slgr-2015-0022.

Korbak, T. (2019a). Computational enactivism under the free energy principle. *Synthese*, first online, 1–21. https://doi.org/10.1007/s11229-019-02243-4.

Korbak, T. (2019b). Unsupervised learning and the natural origins of content. *Avant*, 3(10), 1–14. https://doi.org/10.26913/avant.2019.03.11.

Korsgaard, Ch. (1996). *The sources of normativity*. New York: Cambridge University Press.

Korsgaard, Ch. (2009). *Self-constitution: Agency, identity, and integrity*. Oxford: Oxford University Press.

Kretzmer, E. R. (1952). Statistics of television signals. *Bell System Technical Journal*, 4(31), 751–763.

Kripke, S. (1982). *Wittgenstein on rules and private language*. Cambridge: Harvard University Press.

Krohs, H., Kroes, P. (eds.) (2009). *Functions in biological and artificial worlds*. Cambridge: MIT Press.

Krokos, J. (2006). Nastawienie i jego rola poznawcza. Zarys problemu. W: S. Koperek, W. Zuziak (red.), *Scire Deum* (281–291). Kraków: WN PAT.

Kwisthout, J., Bekkering, H., van Rooij, I. (2017). To be precise, the details don't matter: On predictive processing, precision, and level of detail of predictions. *Brain and Cognition*, 112, 84–91.

Kwisthout, J., van Rooij, I. (2013). Bridging the gap between theory and practice of approximate Bayesian inference. *Cognitive Systems Research*, 24, 2–8.

Kwisthout, J., van Rooij, I. (2019). Computational resource demands of a predictive Bayesian brain. *Synthese*, first online, 1–15. https://doi.org/10.1007/s42113-019-00032-3.

Latawiec, A. (1992). Wokół pojęcia zjawiska biologicznego. *Studia Philosophiae Christianae*, 2(28), 241–254.

Lau, J., Deutsch, M. (2002). Externalism about mental content. W: E. N. Zalta (red.), *Stanford encyclopedia of philosophy* (1–13). Pobrano z: https://plato.stanford.edu/archives/fall2019/entries/content-externalism/ (dostęp: 2.08.2019).

Laudan, L. (1977). *Progress and its problems: Towards a theory of scientific growth*. Berkeley: University of California Press.

Lee, T. S., Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Optical Society of America*, 20(7), 1434–1448.

Leszczyński, D. (2010). *Struktura poznawcza i obraz świata. Zagadnienie podmiotowych warunków poznania we współczesnej filozofii*. Wrocław: Wydawnictwo UWr.

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, 4(61), 343–377.

Levins, R. (1966). The strategy of model building in population biology. *American scientist*, 4(54), 421–431.

Little, M. (1994). Moral realism II: Non-naturalism. *Philosophical Books*, 35, 225–232.

Litwin, P., Miłkowski, M. (2020). Unification by fiat: Arrested development of predictive processing. *Cognitive Science*, 7(44), 1–27. https://doi.org/10.1111/cogs.12867.

Longo, G., Montévil, M. (2013). *Perspectives on organisms: Biological time, symmetries and singularities*. Berlin: Springer.

Longo, G., Montévil, M., Kauffman, S. (2012). No entailing laws, but enablement in the evolution of the biosphere. W: Ch. Igel, D. V. Arnold (eds.), *Proceedings of the 14th International Conference on Genetic and Evolutionary Computation ConferenceCompanion* (1379–1392). New York: ACM.

Luciw, M., Kompella, V., Kazerounian, S., Schmidhuber, J. (2013). An intrinsic value system for developing multiple invariant representations with incremental slowness learning. *Frontiers in Neurorobotics*, 7(9), 1–19. https://doi.org/10.3389/fnbot.2013.00009.

Lukac, M., Kameyama, M. (2015). An algorithm selection based platform for image understanding using high-level symbolic feedback and machine learning. *International Journal of Machine Learning and Cybernetics*, 6, 417–434. https://doi.org/10.1007/s13042-013-0197-x.

Lyons, T. D. (2005). Towards a purely axiological scientific realism. *Erkenntnis*, 63(2), 167–204. https://doi.org/10.1007/s10670-005-3225-8.

Lyu, S., Simoncelli, E. P. (2007). Statistically and perceptually motivated nonlinear image representation, human vision and electronic imaging. W: *XII, Proc. Of SPIE-IS&T Electronic Imaging* (1–15). SPIE.

Machamer, P., Darden, L. Craver, C. F. (2011). Myślenie w kategoriach mechanizmów (W. Hensel, tłum.). *Przegląd Filozoficzno-Literacki*, 2–3(31), 145–176.

Maciejczak, M. (2005). Percepcyjne normowanie i efekt prototypowy. *Przegląd Filozoficzny – Nowa Seria*, 1(53), 185–195.

Maciejczak, M. (2007). *Świadomość i sens*. Warszawa: Wydawnictwo Naukowe UKSW.

Maciejczak, M. (2008). Pojęcie modelu świata. *Studia Philosophiae Christianae*, 2(44), 51–54.

MacKay, D. M. (1956). The epistemological problem for automata. W: C. E. Shannon, J. McCarthy (eds.), *Automata studies* (235–251). Princeton: Princeton University Press.

Maher, Ch. (2017). *Plant minds: A philosophical defense*. Oxford: Routledge.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proc. IEEE*, 63, 561–580.

Marr, D. (1982). *Vision: A computational approach*. San Francisco: Freeman & Co.

Marr, D., Poggio, T. (1976). Cooperative computation of stereo disparity. *Science*, 194, 283–287.

Marr, D., Poggio, T. (1979). A computational theory of human stereo vision. *Proceedings of the Royal Society of London B*, 204, 301–328.

Maynard-Smith, J. (2000). The concept of information in biology. *Philosophy of Science*, 67(2), 689–699. https://doi.org/10.1086/392768.

Mayr, E. (2002). *To jest biologia* (J. Szacki, tłum.). Warszawa: Prószyński i S-ka.

McAllister, J. W. (2007). Model selection and the multiplicity of patterns in empirical data. *Philosophy of Science*, 74, 884–894.

McAllister, J. W. (2009). What do patterns in empirical data tell us about the structure of the world? *Synthese*, 182, 73–87. https://doi.org/10.1007/s11229-009-9613-x.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, 14(8), 348–356.

McDowell, J. (2013). Perceptual experience: Both relational and contentful. *European Journal of Philosophy*, 21(1), 144–157.

McGregor, S. (2017). The Bayesian stance: Equations for 'as-if' sensorimotor agency. *Adaptive Behavior*, 2(25), 72–82. https://doi.org/10.1177/1059712317700501.

McKay, R. T., Dennett, D. (2009). The evolution of misbelief. *Behavioral and Brain Sciences*, 32, 493–510.

Miller, G. A., Galanter, E., Pribram, K. (1980). *Plany i struktura zachowania*. Warszawa: Wydawnictwo Naukowe PWN.

Miller, P. (2016). Dynamical systems, attractors, and neural circuits. *F1000Res*. 5. https://doi.org/10.12688/f1000research.7698.1.

Millikan, R. G. (1984). *Language, thought, and other biological categories: New foundations for realism*. Cambridge: MIT Press.

Millikan, R. G. (1989). In defense of proper functions. *Philosophy of Science*, 2(56), 288–302.

Millikan, R. G. (2011). Funkcje biologiczne – dwa paradygmaty (J. Klimczyk, tłum.). *Przegląd Filozoficzno-Literacki*, 2–3(31), 107–143.

Miłkowski, M. (2004). Filozofia jako inżynieria odwrotna: Rzecz o filozofii Daniela Dennetta. *Przegląd Filozoficzny – Nowa Seria*, 2(50), 75–89.

Miłkowski, M. (2007). Siemek, dialektyka, rzeczywistość. O transcendentalnej filozofii społecznej. *Przegląd Filozoficzno-Literacki*, 1(16), 99–113.

Miłkowski, M. (2013). Reverse engineering in cognitive science. W: M. Miłkowski, K. Talmont-Kamiński (eds.), *Regarding the mind, naturally: Naturalist approaches to the sciences of the mental* (12–29). Newcastle upon Tyne: Cambridge Scholars Publishing.

Miłkowski, M. (2015a). Satisfaction conditions in anticipatory mechanisms. *Biology & Philosophy*, 30(5), 709–728. https://doi.org/10.1007/s10539-015-9481-3.

Miłkowski, M. (2015b). The hard problem of content: Solved (long ago). *Studies in Logic, Grammar and Rhetoric*, 41(1), 73–88. https://doi.org/10.1515/slgr-2015-0021.

Miłkowski, M. (2016a). Integrating cognitive (neuro)science using mechanisms. *Avant*, 6(2), 45–67. https://doi.org/10.26913/70202016.0112.0003.

Miłkowski, M. (2016b). Unification strategies in cognitive science. *Studies in Logic, Grammar and Rhetoric*, 48(1), 13–33. https://doi.org/10.1515/slgr-2016-0053.

Miłkowski, M. (2018a). Autopoiesis nie wywołała rewolucji. W: G. Króliczak, K. Łastowski, Ł. Przybylski, P. Przybysz, M. Urbański (red.), *Filozof w krainie umysłów. Profesorowi Andrzejowi Klawiterowi w darze* (217–227). Poznań: Wydawnictwo Naukowe Wydziału Nauk Społecznych UAM.

Miłkowski, M. (2018b). Morphological computation: Nothing but physical computation. *Entropy*, 20(12), 1–18. https://doi.org/10.3390/e20120942.

Miłkowski, M., Clowes, R., Rucińska, Z., Przegalińska, A., Zawidzki, T., Krueger, J., Gies, A., McGann, M., Afeltowicz, Ł., Wachowski, W., Stjernberg, F., Loughlin, V., Hohol, M. (2018). From wide cognition to mechanisms: A silent revolution. *Frontiers in Psychology*, 9(2393), 1–17. https://doi.org/10.3389/fpsyg.2018.02393.

Mitchell, S. (2003). *Biological complexity and integrative pluralism*. Cambridge: Cambridge University Press.

Montero, B., Papineau, D. (2005). A defence of the via negativa argument for physicalism. *Analysis*, 65, 233–237.

Moń, R. (2011). *Warto czy należy? Studium na temat istoty i źródeł normatywności*. Warszawa: Wydawnictwo Naukowe UKSW.

Moore, G. E. (1903). *Principia ethica*. New York: Cambridge University Press.

Moreno, A. (2016). Some conceptual issues in the transition from chemistry to biology. *History & Philosophy of the Life Sciences,* 38(4), 1–19. https://doi.org/10.1007/s40656-016-0117-y.

Mortimer, H. (1982). *Logika indukcji*. Warszawa: Wydawnictwo Naukowe PWN.

Morville, T., Friston, K., Burdakov, D., Siebner, H. R., Hulme, O. J. (2018). *The homeostatic logic of reward* (1–43). bioRxiv, 242974. https://doi.org/10.1101/242974.

Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R. J., Friston, K. J. (2014b). Bayesian inferences about the self (and others): A review. *Consciousness and Cognition*, 25(100), 67–76. https://doi.org/10.1016/j.concog.2014.01.009.

Moutoussis, M., Trujillo-Barreto, N. J., El-Deredy, W., Dolan, R. J., Friston, K. J. (2014a). A formal model of interpersonal inference. *Frontiers in Human Neuroscience*, 8:160, 1–12. https://doi.org/10.3389/fnhum.2014.00160.

Muders, S. (2014). On the concept of the normative in the assessment of mental disorder. *Frontiers in Psychology*, 5:129, 1–3. https://doi.org/10.3389/fpsyg.2014.00129.

Müller, A., Müller, K. (eds.) (2007). *An unfinished revolution*. Austria: Edition Echoraum.

Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics*, 66(3), 241–251.

Nair, V., Susskind, J., Hinton, G. E. (2008). Analysis-by-synthesis by learning to invert generative black boxes. W: V. Kůrková, R. Neruda, J. Koutník (eds.), *Artificial neural networks – ICANN 2008. ICANN 2008. Lecture notes in computer science*, Vol. 5163 (1–10). Berlin, Heidelberg: Springer.

Neander, K. (1995). Misrepresenting & malfunctioning. *Philosophical Studies*, 2(79), 109–141.

Neisser, U. (1967). *Cognitive psychology*. Englewood Cliffs: Prentice Hall.

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. New York: Freeman.

Ng, Y. N., Jordan, M. I. (2001). On discriminative vs. generative classifiers: Acomparison of logistic regression and naive Bayes. *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, 841–848. [b.m.]: [b.w.].

Noë, A. (2004). *Action in perception*. Cambridge: MIT Press.

Norman, D. A. (2018). *Dizajn na co dzień* (D. Malina, tłum.). Kraków: Wydawnictwo Karakter.

Nowakowski, P. (2018). Co wyjaśniają modele Fristona? Referat wygłoszony 6.03.2018 r. podczas warsztatów *Przetwarzanie predykcyjne – możliwości i ograniczenia*, zorganizowanych w IF UKSW (praca niepublikowana).

Oaksford, M. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, 5(8), 349–357.

Oaksford, M. (2014). Normativity, interpretation and Bayesian models. *Frontiers in Psychology*, 5(332), 1–5. https://doi.org/10.3389/fpsyg.2014.00332.

Oaksford, M., Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.

O'Brien, G., Opie, J. (2004). Notes towards a structuralist theory of mental representation. W: H. Clapin, P. Staines, P. Slezak (eds.), *Representation in mind: New approaches to mental representation* (1–20). Amsterdam: Elsevier.

Ofner, A., Stober, S. (2018). *Hybrid active inference* (preprint). arXiv:1810.02647v1.

Oliver, B. (1952). Efficient coding. *Bell System Technical Journal*, 31(4), 724–750.

O'Regan, J. K., Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(05), 939–973.

Orlandi, N. (2016). Bayesian perception is ecological perception. *Philosophical Topics*, 44(2), 327–351.

Orlandi, N. (2017). Predictive perceptual systems. *Synthese*, first online, 1–20. https://doi.org/10.1007/s11229-017-1373-4.

Orlandi, N., Lee, G. (2018). How radical is predictive processing? W: M. Colombo, E. Irvine, M. Stapleton (eds.), *Andy Clark & critics* (1–14). Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780190662813.003.0016.

Papineau, D. (2007). Naturalism. W: E. N. Zalta (ed.), *The Stanford encyclopedia of philosophy* (Summer 2020 Edition), Pobrano z: https://plato.stanford.edu/archives/sum2020/entries/naturalism/ (dostęp: 17.06.2020).

Park, H. J., Friston, K. J. (2013). Structural and functional brain networks: From connections to cognition. *Science*, 342, 579–589. https://doi.org/10.1126/science.1238411.

Parpart, P., Jones, M., Love, B. C. (2018). Heuristics as Bayesian inference under extreme priors. *Cognitive Psychology*, 102, 127–144. https://doi.org/10.1016/j.cogpsych.2017.11.006.

Parr, T., Friston, K. J. (2018). The anatomy of inference: Generative models and brain structure. *Frontiers in Computational Neuroscience*, 12(90), 1–23. https://doi.org/10.3389/fncom.2018.00090.

Pattee, H. H. (1968). The physical basis of coding and reliability in biological evolution. W: C. H. Waddington (ed.), *Towards a theoretical biology*, Vol. 1 (33–54). Edinburgh: Edinburgh University Press.

Pattee, H. H. (1972). Laws and constraints, symbols and languages. W: C. H. Waddington (ed.), *Towards a theoretical biology*, Vol. 4 (248–258). Edinburgh: Edinburgh University Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco: Morgan Kaufmann Publishers.

Pezzulo, G. (2017). Tracing the roots of cognition in predictive processing. W: T. Metzinger, W. Wiese (eds.), *Philosophy and Predictive Processing*, 20, 1–20. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573215.

Pezzulo, G., Cartoni, E., Rigoli, F., Pio-Lopez, L., Friston, K. J. (2016). Active inference, epistemic value, and vicarious trial and error. *Learning & Memory*, 23(7), 322–338.

Pezzulo, G., Rigoli, F., Friston, K. J. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology*, 134, 17–35.

Pfeifer, R., Bongard, J. (2006). *How the body shapes the way we think: A new view of intelligence*. Cambridge: MIT Press.

Piccinini, G., Craver, C. F. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, 183(3), 283–311.

Pickering, M. J., Clark, A. (2014). Getting ahead: Forward models and their place in cognitive architecture. *Trends in Cognitive Science*, 8(9), 451–456. https://doi.org/10.1016/j.tics.2014.05.006.

Piekarski, M. (2016). Od typiki doświadczenia do normatywnej antycypacji. Przyczynek do fenomenologii normatywności. *Filo-Sofija*, 2(33), 71–86.

Piekarski, M. (2017a). Commentary: Getting into predictive processing's great guessing game: Bootstrap heaven or hell? *Frontiers in Psychology*, 8(1244), 1–3. https://doi.org/10.3389/fpsyg.2017.01244.

Piekarski, M. (2017b). Commentary: Brain, mind, world: Predictive coding, Neo-Kantianism, and transcendental idealism. *Frontiers in Psychology*, 8(2077), 1–3. https://doi.org/10.3389/fpsyg.2017.02077.

Piekarski, M. (2018). One or many normativities? *Studia Philosophiae Christianae*, 1(54), 6–24. https://doi.org/10.21697/2018.54.1.11.

Piekarski, M. (2019). Normativity of predictions: A new research perspective. *Frontiers in Psychology*, 10(1710), 1–4. https://doi.org/10.3389/fpsyg.2019.01710.

Piekarski, M. (2020). Normatywne mechanizmy reprezentacyjne a koherencyjna koncepcja błędu reprezentacyjnego. *Avant*, 3(11), 1–21. https://doi.org/10.26913/avant.2020.03.11.

Piekarski, M., Wachowski, W. (2018). Artefacts as social things: Design-based approach to normativity. *Techné*, 22, 400–424. https://doi.org/10.5840/techne2018121990.

Piłat, R. (1999). *Umysł jako model świata*. Warszawa: Wydawnictwo IFIS PAN.

Poczobut, R. (2009). *Między redukcją a emergencją. Spór o miejsce umysłu w świecie fizycznym*. Wrocław: Wydawnictwo UWr.

Poczobut, R. (2011). Od rozszerzonego życia do rozszerzonego umysłu (poznania). *Przegląd Filozoficzno-Literacki*, 2–3(31), 217–238.

Povich, M. (2019). Model-based cognitive neuroscience: Multifield mechanistic integration in practice. *Theory & Psychology*, 29(5), 640–656. https://doi.org/10.1177/0959354319863880.

Półtawski, A. (1973). *Świat, spostrzeżenie, świadomość*. Warszawa: Wydawnictwo Naukowe PWN.

Półtawski, A. (1986). Poznanie i zmysły. *Studia Filozoficzne*, 1–2, 97–105.

Protevi, J. (2006). Enaction. W: J. Protevi (ed.), *A dictionary of continental philosophy* (169–170). Yale: Yale University Press.

Quine, W. V. O. (1986). *Granice wiedzy i inne eseje filozoficzne* (B. Stanosz, tłum.). Warszawa: PIW.

Ramsey, W. M. (2007). *Representation reconsidered*. Cambridge: Cambridge University Press.

Ramsey, W. M. (2016). Untangling two questions about mental representation. *New Ideas Psychology*, 40, 3–12. https://doi.org/10.1016/j.newideapsych.2015.01.004.

Ramstead, M. J. D., Badcock, P. B., Friston, K. J. (2017). Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews*, 24, 1–16. https://doi.org/10.1016/j.plrev.2017.09.001.

Ramstead, M. J. D., Badcock, P. B., Friston, K. J. (2018). Variational neuroethology: Answering further questions: Reply to comments on „answering Schrödinger's question: A free-energy formulation". *Physics of Life Reviews*, 24, 59–66.

Ramstead, M. J. D., Kirchhoff, M., Constant, A., Friston, K. J. (2019). Multiscale integration: Beyond internalism and externalism. *Synthese*, first onlince, 1–30. https://doi.org/10.1007/s11229-019-02115-x.

Ramstead, M. J. D., Kirchhoff, M. D., Friston, K. J. (2019). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior*, first online, 1–15. https://doi.org/10.1177/1059712319862774.

Ramstead, M. J. D., Veissière, S. P. L., Kirmayer, L. J. (2016). Cultural affordances: Scaffolding local worlds through shared intentionality and regimes of attention. *Frontiers in Psychology*, 7(1090), 1–21. https://doi.org/10.3389/fpsyg.2016.01090.

Rao, R. P., Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), 79–87. https://doi.org/10.1038/4580.

Ravenscroft, I. (2019). Folk psychology as a theory. W: E. N. Zalta (ed.), *The Stanford encyclopedia of philosophy* (Summer 2019 Edition). Pobrano z: https://plato.stanford.edu/archives/sum2019/entries/folkpsych-theory/ (dostęp: 28.05.2019).

Rączaszek-Leonardi, J. (2012). Language as a system of replicable constraints. W: H. H. Patee, J. Rączaszek-Leonardi (eds.), *LAWS, LANGUAGE and LIFE. Biosemiotics*, Vol. 7 (295–333). Dordrecht: Springer.

Rescorla, M. (2015). Bayesian perceptual psychology. W: M. Matthen (ed.), *The Oxford handbook of philosophy of perception* (694–716). Oxford: Oxford University Press.

Ridge, M. (2003). Moral non-naturalism. W: E. N. Zalta (ed.), *The Stanford encyclopedia of philosophy* (Fall 2019 Edition). Pobrano z: https://plato.stanford.edu/archives/fall2019/entries/moral-non-naturalism/ (dostęp: 21.06.2020).

Rietveld, E. (2008). Situated normativity: The normative aspect of embodied cognition in unreflective action. *Mind*, 117(468), 973–1001.

Rosati, C. S. (2003). Agency and the open question argument. *Ethics*, 113, 490–527.

Rosenblueth, A., Wiener, N., Bigelow, J. (1943). Behavior, purpose and teleology. *Philosophy of Science*, 10, 18–24.

Rubinov, M., Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *Neuroimage*, 52(3), 1059–1069. https://doi.org/10.1016/j.neuroimage.2009.10.003.

Sady, W. (2013). *Spór o racjonalność naukową. Od Poincarégo do Laudana*. Toruń: Wydawnictwo UMK.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.

Savage, L. J. (1954). *The foundations of statistics*. New York: John Wiley.

Scheler, M. (2014). *Der formalismus in der ethik und die materiale wertethik: Neuer versuch der grundlegung eines ethischen personalismus*. Hamburg: Felix Meiner Verlag.

Schetz, A. (2014). *Biologiczny eksternalizm w teoriach percepcji*. Szczecin: WN USz.

Schwartenbeck, P., FitzGerald, T., Dolan, R. J., Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. *Frontiers in Psychology*, 4(710), 1–5. https://doi.org/10.3389/fpsyg.2013.00710.

Schwartenbeck, P., FitzGerald, T., Mathys, C., Dolan, R., Kronbichler, M., Friston, K. J. (2015). Evidence for surprise minimization over value maximization in choice behavior. *Scientific Reports*, 5, 1–14. https://doi.org/10.1038/srep16575.

Searle, J. (1999). *Umysł na nowo odkryty* (T. Baszniak, tłum.). Warszawa: PIW.

Sellars, W. (1962/1991). Philosophy and the scientific image of man. W: W. Sellars (ed.), *Science, perception and reality* (1–40). Atascadero: Ridgeview Publishing Company.

Seth, A. K. (2015). Inference to the best prediction. W: T. Metzinger, J. M. Windt (eds.), *Open MIND*, 35R, 1–8. Frankfurt am Main: MIND Group.

Seth, A., Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Philosophical Transactions Of The Royal Society B: Biological Sciences*, 371(1708), 1–10. https://doi.org/10.1098/rstb.2016.0007.

Shagrir, O. (2010). Marr on computational-level theories. *Philosophy of Science*, 77(4), 477–500.

Shagrir, O. (2012). Structural representations and the brain. *The British Journal for the Philosophy of Science*, 63, 519–545. https://doi.org/10.1093/bjps/axr038.

Shams, L., Ma, W. J., Beierholm, U. (2005). Sound-induced flash illusion as an optimal percept. *Neuroreport*, 16, 1923–1927. https://doi.org/10.1097/01.wnr.0000187634.68504.bb.

Shannon, C. E., Weaver, W. (1964). *The mathematical theory of communication*. Urbana and Chicago: University of Illinois Press.

Shannon, E. A. (1948). Mathematical theory of communication. *Bell Labs Technical Journal*, 27, 379–423.

Shea, H., Godfrey-Smith, P., Cao, R. (2018). Content in simple signalling systems. *British Journal for the Philosophy of Science*, online first. https://doi.org/10.1093/bjps/axw036.

Shea, N. (2014). Exploitable isomorphism and structural representation. *Proceedings of the Aristotelian Society*, CXIV, 77–92. https://doi.org/10.1111/j.1467-9264.2014.00367.x.

Shi, Y. Q., Sun, H. (1999). *Image and video compression for multimedia engineering: Fundamentals, algorithms, and standards*. Boca Raton: CRC Press.

Simon, H. (1983). *Reason in human affairs*. Stanford: Stanford University Press.

Sims, A. (2017). The problems with prediction – the dark room problem and the scope dispute. W: T. Metzinger, W. Wiese (eds.), *Philosophy and Predictive Processing*, 2, 1–18. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573246.

Skyrms, B. (2010). *Signals: Evolution, learning, & information*. Oxford–New York: Oxford University Press.

Sober, E. (1984). *The nature of selection: Evolutionary theory in philosophical focus*. Chicago: University of Chicago Press.

Solopchuk, O. (2018). *Tutorial on active inference*. Pobrano z: https://medium.com/@solopchuk/tutorial-on-active-inference-30edcf50f5dc (dostęp: 27.08.2019).

Sporns, O. (2014). Contributions and challenges for network models in cognitive neuroscience. *Nature Neuroscience*, 17, 652–660. https://doi.org/10.1038/nn.3690.

Spratling, M. W. (2012). Unsupervised learning of generative and discriminative weights encoding elementary image components in a predictive coding model of cortical function. *Neural Computation*, 24(1), 60–103. https://doi.org/10.1162/NECO_a_00222.

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition*, 112, 92–97. https://doi.org/10.1016/j.bandc.2015.11.003.

Sprevak, M. (2013). Fictionalism about neural representations. *Monist*, 96, 539–560. https://doi.org/10.5840/monist201396425.

Srinivasan, M. V., Laughlin, S. B., Dubs, A. (1982). Predictive coding: A fresh view of inhibition in the retina. *Proceedings of the Royal Society of London B: Biological Sciences*, 216(1205), 427–459.

Stanford, P. K. (2019). Unconceived alternatives and conservatism in science: The impact of professionalization, peer-review and big science. *Synthese*, 196(10), 3915–3932. https://doi.org/10.1007/s11229-015-0856-4.

Sterelny, K. Griffiths, P. E. (1999). *Sex and death: An introduction to the philosophy of biology*. Chicago: University of Chicago Press.

Sterelny, K., Smith, K. C., Dickison, M. (1996). The extended replicator. *Biology and Philosophy*, 11(3), 377–403. https://doi.org/10.1007/BF00128788.

Sterling, P., Eyer, J. (1988). Allostasis: A new paradigm to explain arousal pathology. W: S. Fisher, J. Reason (eds.), *Handbook of life stress, cognition, and health* (629–649). Chichester: John Wiley & Sons.

Stern, R. (2019). Transcendental arguments. W: E. N. Zalta (ed.), *The Stanford encyclopedia of philosophy* (Summer 2019 Edition). Pobrano z: https://plato.stanford.edu/archives/sum2019/entries/transcendental-arguments/ (dostęp: 15.02.2019).

Stier, M. (2013). Normative preconditions for the assessment of mental disorder. *Frontiers in Psychology*, 4(611), 1–9. https://doi.org/10.3389/fpsyg.2013.00611.

Ströing, P. (2018). *Scientific phenomena and patterns in data*. Dissertation, LMU München: Faculty of Philosophy, Philosophy of Science and the Study of Religion.

Sun, Z., Firestone, Ch. (2020). The dark room problem. *Trends in Cognitive Sciences*, 5(24), 346–348. https://doi.org/10.1016/j.tics.2020.02.006.

Suñé, A., Martínez, M. (2019). Real patterns and indispensability. *Synthese*, first online, 1–16. https://doi.org/10.1007/s11229-019-02343-1.

Swanson, L. R. (2016). The predictive processing paradigm has roots in Kant. *Frontiers in Systems Neuroscience*, 10(79), 1–13. https://doi.org/10.3389/fnsys.2016.00079.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022), 1279–1285.

Thagard, P. (2009). Why cognitive science needs philosophy and vice versa. *Topics in Cognitive Science*, 1(2), 237–254. https://doi.org/10.1111/j.1756-8765.2009.01016.x.

Thagard, P., Verbeurgt, K. (1998). Coherence as constraint satisfaction. *Cognitive Science*, 1(22), 1–24. https://doi.org/10.1207/s15516709cog2201_1.

Thompson, E. (2011). Précis of mind in life. *Journal of Consciousness Studies*, 18, 10–22.

Thornton, C. (2016). *Predictive processing is Turing complete: A new view of computation in the brain*. Pobrano z: http://users.sussex.ac.uk/~christ/drafts/draft-o.pdf (dostęp: 10.02.2020).

Tomasello, M. (2015). *Dlaczego współpracujemy?* (Ł. Kwiatek, tłum.). Kraków: Copernicus Center Press.

Turing, A. M. (1968). Intelligent machinery. W: C. R. Evans, A. D. F. Robertson (eds.), *Cybernetics*, 107–127. Philadelphia: University City Press.

Tyszka, T. (2010). *Decyzje: Perspektywa psychologiczna i ekonomiczna*. Warszawa: Wydawnictwo Scholar.

Tyszka, T., Zaleśkiewicz, T. (2001). *Racjonalność decyzji. Pewność i ryzyko*. Warszawa: Polskie Wydawnictwo Ekonomiczne.

Umerez, J., Mossio, M. (2013). Constraint. W: W. Dubitzky, O. Wolkenhauer, K. H. Cho, H. Yokota (eds.), *Encyclopedia of systems biology* (490–493). Berlin: Springer. https://doi.org/10.1007/978-1-4419-9863-7.

Wachowski, W. M. (2018a). Normatywność usytuowana. Ujęcie ekologiczne. *Studia Philosophiae Christianae*, 1(54), 143–165. https://doi.org/10.21697/2018.54.1.16.

Wachowski, W. M. (2018b). Commentary: Distributed cognition and distributed morality: Agency, artifacts and systems. *Frontiers in Psychology*, 9(490), 1–3. https://doi.org/10.3389/fpsyg.2018.00490.

Waleszczyński, A. (2018). Efekt Knobe'a z perspektywy etyki tomistycznej. Problem porządków i kompetencji normatywnych. *Studia Philosophiae Christianae*, 1(54), 65–92. https://doi.org/10.21697/2018.54.1.13.

Walker, R. C. S. (2006). Kant and transcendental arguments. W: P. Guyer (ed.), *Kant and modern philosophy*. Cambridge: Cambridge University Press.

Wallace, D. (2003). Everett and structure. *Studies in History and Philosophy of Modern Physics*, 34, 87–105.

Ward, D., Silverman, D., Villalobos, M. (2017). Introduction: The varieties of enactivism. *Topoi*, 36, 365–375. https://doi.org/10.1007/s11245-017-9484-6.

Warren, W. H. (1984). Perceiving affordances: Visual guidance of stair climbing. *Journal of Experimental Psychology: Human Perception and Performance*, 10(5), 683–703.

Weiskopf, D. A. (2016). Integrative modeling and the role of neural constraints. *Philosophy of Science*, 83, 674–685. https://doi.org/10.1086/687854.

Whiting, D. (2013). What is the normativity of meaning? *Inquiry*, 3(59), 219–238.

Wiese, W. (2016a). What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences*, 16, 715–736. https://doi.org/10.1007/s11097-016- 9472-0.

Wiese, W. (2016b). Action is enabled by systematic misrepresentations. *Erkenntnis*, 82, 1233–1252.

Wiese, W., Metzinger, T. (2017). Vanilla PP for philosophers: A primer on predictive processing. W: T. Metzinger, W. Wiese (eds.), *Philosophy and Predictive Processing*, 1, 1–18. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573024.

Williams, D. (2018). Predictive coding and thought. *Synthese*, first online, 1–27. https://doi.org/10.1007/s11229-018-1768-x.

Williams, D. (2019). Epistemic irrationality in the Bayesian brain. *British Journal for the Philosophy of Science*, 1-21. https://doi.org/10.1093/bjps/axz044

Winning, J. (2018). Mechanistic causation and constraints: Perspectival parts and powers, non-perspectival modal patterns. *The British Journal for the Philosophy of Science*, *axy042*. https://doi.org/10.1093/bjps/axy042.

Winning, J., Bechtel, W. (2018). Rethinking causality in biological and neural mechanisms: Constraints and control. *Minds and Machines*, 2(28), 287–310. https://doi.org/10.1007/s11023-018-9458-5.

Woleński, J. (2001). Naturalizm, antynaturalizm i podstawy statystyki. *Filozofia Nauki*, 9(1), 147–155.

Woodward, J. (2009). Data and phenomena: A restatement and defense. *Synthese*, 182, 165–179. https://doi.org/10.1007/s11229-009-9618-5.

Wouters, A. G. (2003). Four notions of biological function. *Studies in History and Philosophy of Biology and Biomedical Science*, 34(4), 633–668.

Wright, L. (1973). Functions. *The Philosophical Review*, 2(82), 139–168.

Van der Cruys, S., Friston, K. J., Clark, A. (w przygotowaniu). Controlled optimism: Reply to sun and Firestone on the dark room problem. *Trends in Cognitive Sciences*.

Van Es, T. (2020). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior*, first online, 1–15. https://doi.org/10.1177/1059712320918678.

van Fraassen, C., Bas, C. (1980). *The scientific image*. Oxford: Oxford University Press.

Van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy*, 92(7), 345–381.

van Rooij, I., Wright, C. D., Kwisthout, J., Wareham, T. (2018). Rational analysis, intractability and the prospects of 'as if'-explanations. *Synthese*, 195(2), 491–510. https://doi.org/10.1007/s11229-014-0532-0.

Varela, F. (1997). Patterns of life: Intertwining identity and cognition. *Brain and Cognition*, 34(1), 72–87.

Varela, F., Maturana, H., Uribe, R. (1974). Autopoiesis: The organization of living systems, its characterization and a model. *Biosystems*, 5(4), 187–196.

Varela, F., Thompson, E., Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, London: The MIT Press.

Veissière, S., Constant, A., Ramstead, M. J. D., Firston, K. J., Kirmayer, L. J. (2020). Thinking through other minds: A variational approach to cognition and culture. *Behavioral and Brain Sciences*, 43, 1–97. https://doi.org/10.1017/S0140525X19001213.

Velleman, J. D. (2000). *The possibility of practical reason*. Oxford: Oxford University Press.

Villalobos, M. (2013). Enactive cognitive science: Revisionism or revolution? *Adaptive Behavior*, 3(21), 159–167. https://doi.org/10.1177/1059712313482953.

Villalobos, M., Dewhurst, J. (2018). Enactive autonomy in computational systems. *Synthese*, 195(5), 1891–1908. https://doi.org/10.1007/s11229-017-1386-z.

Vogel, J. (1990). Cartesian skepticism and inference to the best explanation. *The Journal of Philosophy*, 87(11), 658–666.

Ylikoski, P. (2013). Causal and constitutive explanation compared. *Erkenntnis*, 78, 277–297. https://doi.org/10.1007/s10670-013-9513-9.

Ylikoski, P. (2015). Social mechanism. W: J. D. Wright (ed.), *International encyclopedia of the social & behavioral sciences*, Vol. 22 (415–420), 2nd ed. Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.03194-9.

Zahavi, D. (2018). Brain, mind, world: Predictive coding, neo-Kantianism, and transcendental idealism. *Husserl Studies*, 34, 47–61. https://doi.org/10.1007/s10743-017-9218-z.

Zednik, C. (2008). Dynamical models and mechanistic explanations. W: B. C. Love, K. McRae, V. M. Sloutsky (eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (1454–1459). Austin: Cognitive Science Society.

Zednik, C. (2017). Mechanisms in cognitive science. W: S. Glennan, P. Illari (eds.), *The Routledge handbook of mechanisms and mechanical philosophy* (389–400). London: Routledge, Taylor & Francis Group.

Zednik, C., Jäkel, F. (2014). How does Bayesian reverse-engineering work? W: M. Bello, M. Guarini, M. McShane, B. Scassellati (eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (666–671). Austin: Cognitive Science Society.

Zednik, C., Jäkel, F. (2016). Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese*, 193(12), 3951–3985. https://doi.org/10.1007/s11229-016-1180-3.

Predictive mechanisms and their normativity

Summary

The aim of this study is to justify the belief that there are biological normative mechanisms that fulfill non-trivial causal roles in the explanations (as formulated by researchers) of actions and behaviors present in specific systems. One example of such mechanisms is the predictive mechanisms described and explained by predictive processing (hereinafter PP), which (1) guide actions and (2) shape causal transitions between states that have specific content and fulfillment conditions (e.g. mental states). Therefore, I am guided by a specific theoretical goal associated with the need to indicate those conditions that should be met by the non-trivial theory of normative mechanisms and the specific models proposed by PP supporters.

In this work, I use classical philosophical methods, such as conceptual analysis and critical reflection. I also analyze selected studies in the field of cognitive science, cognitive psychology, neurology, information theory and biology in terms of the methodology, argumentation and language used, in accordance with their theoretical importance for the issues discussed in this study. In this sense, the research presented here is interdisciplinary. My research framework is informed by the mechanistic model of explanation, which defines the necessary and sufficient conditions for explaining a given phenomenon. The research methods I chose are therefore related to the problems that I intend to solve.

In the introductory chapter, "The concept of predictive processing", I discuss the nature of PP as well as its main assumptions and theses. I also highlight the key concepts and distinctions for this research framework. Many authors argue that PP is a contemporary version of Kantianism and is exposed to objections similar to those made against the approach of Immanuel Kant. I discuss this thesis and show that it is only in a very general sense that the PP framework is neo-Kantian. Here we are not dealing with transcendental deduction nor with the application of transcendental arguments. I argue that PP is based on reverse engineering and abduction inferences. In the second part of this chapter, I respond to the objection formulated by Dan Zahavi, who directly accuses this research framework of anti-realistic consequences. I demonstrate that the position of internalism, present in the so-called conservative PP, does not imply anti-realism, and that, due to the explanatory role played in it by structural representations directed at real patterns, it is justified to claim that PP is realistic. In this way, I show that PP is a non-trivial research framework, having its subject, specific methods and its own epistemic status. Finally, I discuss positions classified as the so-called radical PP.

In the chapter "Predictive processing as a Bayesian explanatory model" I justify the thesis according to which PP offers Bayesian modeling. Many researchers claim that the brain is an implemented statistical probabilistic network that is an approximation of the

Bayesian rule. In practice, this means that all cognitive processes are to apply Bayes' rule and can be described in terms of probability distributions. Such a solution arouses objections among many researchers and is the subject of wide criticism. The purpose of this chapter is to justify the thesis that Bayesian PP is a non-trivial research framework. For this purpose, I argue that it explains certain phenomena not only at the computational level described by David Marr, but also at the level of algorithms and implementation. Later in this chapter I demonstrate that PP is normative modeling. Proponents of the use of Bayesian models in psychology or decision theory argue that they are normative because they allow the formulation of formal rules of action that show what needs to be done to make a given action optimal. Critics of this approach emphasize that such thinking about the normativity of Bayesian modeling is unjustified and that science should shift from prescriptive to descriptive positions. In a polemic with Shira Elqayam and Jonathan Evans (2011), I show that the division they propose into prescriptivism and Bayesian descriptivism is apparent, because, as I argue, there are two forms of prescriptivism, i.e. the weak and the strong. I argue that the weak version is epistemic and can lead to anti-realism, while the strong version is ontic and allows one to justify realism in relation to Bayesian models. I argue that a weak version of prescriptivism is valid for PP. It allows us to adopt anti-realism in relation to PP. In practice, this means that you can explain phenomena using Bayes' rule. This does not, however, imply that they are Bayesian in nature. However, the full justification of realism in relation to the Bayesian PP presupposes the adoption of strong prescriptivism. This position assumes that phenomena are explained by Bayesian rule because they are Bayesian as such. If they are Bayesian in nature, then they should be explained using Bayesian modeling. This thesis will be substantiated in the chapters "Normative functions and mechanisms in the context of predictive processing" and "Normative mechanisms and actions in predictive processing".

In the chapter "The Free Energy Principle in predictive processing", I discuss the Free Energy Principle (hereinafter FEP) formulated by Karl Friston and some of its implications. According to this principle, all biological systems (defined in terms of Markov blankets) minimize the free energy of their internal states in order to maintain homeostasis. Some researchers believe that PP is a special case of applying this principle to cognition, and that predictive mechanisms are homeostatic mechanisms that minimize free energy. The discussion of FEP is important due to the fact that some authors consider it to be important for explanatory purposes and normative. If this is the case, then FEP turns out to be crucial in explaining normative predictive mechanisms and, in general, any normative biological mechanisms. To define the explanatory possibilities of this principle, I refer to the discussion of its supporters on the issue they define as the problem of continuity and discontinuity between life and mind. A critical analysis of this discussion and the additional arguments I have formulated have allowed me to revise the explanatory ambitions of FEP. I also reject the belief that this principle is necessary to explain the nature of predictive mechanisms. I argue that the principle formulated and defended by Friston is an important research heuristic for PP analysis.

In the chapter "Normative functions and mechanisms in predictive processing", I start my analyzes by formulating an answer to the question about the normative nature of homeostatic mechanisms. I demonstrate that predictive mechanisms are not homeostatic.

I defend the view that a full explanation of normative mechanisms presupposes an explanation of normative functions. I discuss the most important proposals for understanding the normativity of a function, both from a systemic and teleosemantic perspective. I conclude that the non-trivial concept of a function must meet two requirements which I define as explanatory and normative. I show that none of the theories I have invoked satisfactorily meets both of these requirements. Instead, I propose a model of normativity based on Bickhard's account, but supplemented by a mechanistic perspective. I argue that a function is normative when: (1) it allows one to explain the dysfunction of a given mechanism; (2) it contributes to the maintenance of the organism's stability by shaping and limiting possible relations, processes and behaviors of a given system; and when (3) (according to the representational and predictive functions) it enables explaining the attribution of logical values of certain representations / predictions. In such an approach, a mechanism is normative when it performs certain normative functions and when it is constitutive for a specific action or behavior, despite the fact that for some reason it cannot realize it either currently or in the long-term.

Such an understanding of the normativity of mechanisms presupposes the acceptance of the epistemic hypothesis. I argue that this hypothesis is not cognitively satisfactory, and therefore the ontic hypothesis should be justified, which is directly related to adopting the position of ontic prescriptivism. For this reason, referring to the mechanistic theory of scientific explanations, I formulate an ontical interpretation of the concept of a normative mechanism. According to this approach, a mechanism or a function is normative when they perform such and such causal roles in explaining certain actions and behaviors. With regard to the normative properties of predictive mechanisms and functions, this means that they are the causes of specific actions an organism carries out in the environment. In this way, I justify the necessity of accepting the ontic hypothesis and rejecting the epistemic hypothesis.

The fifth chapter, "Normative mechanisms and actions in predictive processing", is devoted to the dark room problem and the related exploration-exploitation trade-off. A dark room is the state that an agent could be in if it minimized the sum of all potential prediction errors. I demonstrate that, in accordance with the basic assumption of PP about the need for continuous and long-term minimization of prediction errors, such a state should be desirable for the agent. Is it really so? Many authors believe it is not. I argue that the test of the value of PP is the possibility of a non-trivial solution of this problem, which can be reduced to the choice between active and uncertainty-increasing exploration and safe and easily predictable exploitation. I show that the solution proposed by PP supporters present in the literature does not enable a fully satisfactory explanation of this dilemma.

Then I defend the position according to which the full explanation of the normative mechanisms, and, subsequently, the solution to the dilemma of exploration and exploitation, involves reference to the existence of constraints present in the environment. The constraints include elements of the environment that make a given mechanism not only causal but also normative. They are therefore key to explaining the predictive mechanisms. They do not only play the role of the context in which the mechanism is implemented, but, above all, are its constitutive component. I argue that the full explanation of the role of

constraints in normative predictive mechanisms presupposes the integration of individual models of specific cognitive phenomena, because only the mechanistic integration of PP with other models allows for a non-trivial explanation of the nature of normative predictive mechanisms that would have a strong explanatory value. The explanatory monism present in many approaches to PP makes it impossible to solve the problem of the dark room.

Later in this chapter, I argue that the Bayesian PP is normative not because it enables the formulation of such and such rules of action, but because the predictive mechanisms themselves are normative. They are normative because they condition the choice of such and such actions by agents. In this way, I justify the hypothesis that normative mechanisms make it possible to explain the phenomenon of agent motivation, which is crucial for solving the dark room problem.

In the last part of the chapter, I formulate the hypothesis of distributed normativity, which assumes that the normative nature of certain mechanisms, functions or objects is determined by the relations into which these mechanisms, functions or objects enter. This means that what is normative (in the primary sense) is the relational structure that constitutes the normativity of specific items included in it. I suggest that this hypothesis opens up many areas of research and makes it possible to rethink many problems.

In the "Conclusion", I summarize the results of my research and indicate further research perspectives.

Indeks nazwisk

(nie obejmuje bibliografii)

A

Abrahamsen, A. 219
Adams, R. A. 26, 42, 178, 210, 215, 252, 253, 257,263.
Adams, W. J. 88,
Aizawa, K. 85
Ajdukiewicz, K. 56, 97
Allen, M. 26, 142, 165, 175, 178, 179, 252
Alston, W. P. 18
Anderson, J. R. 104, 120
Anderson, M. L. 37, 49, 52, 57, 59, 212
Anscombe, G. E. M. 14, 221-223, 224
Anselme, P. 264, 267
Ashby, W. R. 139, 148, 150, 164, 180-182, 283
Audi, R. 16
Austin, J. L. 53

B

Badcock, P. B. 134, 141, 159, 164, 167-168, 246, 264
Ballard, D. H. 36, 43-46, 64, 107,116, 183, 236
Baltieri, M. 237, 244
Barandiaran, X. E. 15, 168
Barwise, J. 270
Bechtel, W. 22-23, 71, 72, 114, 167, 168, 190-191, 219, 227, 269, 279
Beierholm, U. 215
Bekkering, H. 43, 69
Bergstrom, C. T. 156
Bermúdez, J. L. 224
Bernard, C. 135
Bernoulli, D. 249
Bicchieri, C. 14, 15
Bickhard, M. H. 6, 12, 15, 18, 22, 24, 31, 83, 89, 125-126, 155, 177, 186, 193, 201, 203-211, 213-215, 217, 225-228, 266, 304
Bielecka, K. 120, 126, 153, 188, 195, 200, 201-202
Bigelow, J. 181, 183
Bishop, Ch. M. 65
Bock, W. J. 191
Boden, M. A. 149
Bogacz, R. 202, 255
Boghossian, P. 14
Bongard, J. 80
Boorse, C. 187
Bowers, J. S. 104, 110-112, 115
Braillard, P. A. 169
Brandom, R. B. 15, 275
Brooks, R. A. 149
Broome, J. 15
Brown, H. R. 26, 42, 178,
Brożek, B. 11, 14
Bruineberg, J. 25, 60, 79, 82, 86, 137, 139, 143, 214, 247, 255, 292-294
Buckley, Ch. L. 237, 244
Bunke, H. 73
Burge, T. 76, 225, 280
Burr, C. 81, 248

C

Call, J. 20
Cao, R. 156-157
Carnap, R. 97
Cassam, Q. 54
Chater, N. 104, 117, 120
Chemero, A. 26, 37, 49, 52, 57, 82, 86, 140, 149, 161-162, 164, 168, 212, 270-271, 276
Chikofsky, E. J. 113
Chomsky, N. 123-125, 261
Christensen, W. 15, 18-19, 204
Churchland, P. M. 288-289

Churchland, P. S. 15-16
Cisek, P. 81
Clancy, K. 256
Clark, A. 24-25, 36-38, 44, 46, 50, 52, 57, 59, 64, 74-75, 77, 79-81, 92-93, 107, 118, 132-134, 136, 140,143, 148, 152, 163, 166, 174, 197, 210, 212, 238, 239-240, 243, 246-247, 281, 286, 288,
Clark, J. E. 183
Collell, G. 136
Collins, E. 261
Colombo, M. 26, 104, 106, 109-111, 114, 118-119, 122, 168, 170-173, 247, 253, 298, 306,
Conant, R. C. 149, 150, 283
Constant, A. 20, 110, 244-248, 252, 255, 260, 266,
Copp, D. 15-16
Corcoran, A. W. 184
Cosmides, L. 111
Craik, K. H. 137, 182
Craver, C. F. 22, 71-72, 105, 114, 167-168, 172-173, 190-191, 199, 218-221, 226, 229, 268-270, 273, 279
Crisp, R. 16
Cummins, R. 6, 71, 126, 182, 186, 188-189, 191, 193

D
Danks, D. 73, 111, 279
Darden, L. 15, 114, 219-220, 226, 269, 279-280
Daunizeau, J. 165, 167, 211, 246-247, 257
Davidson, D. 14, 86, 221-223
Davies, P. S. 15, 198-201, 205
Davis, C. J. 104, 110-112, 115
de Finetti, B. 98
de Jaegher, H. 84
de Pinedo, M. 289
de Waal, F. 15
de Wit, M. M. 270
Dennett, D. 21, 51, 59, 65-67, 70, 164, 224, 274, 289
Dewhurst, J. 142, 223
Devlin, K. 271
Dickison, M. 153-154
Di Paolo, E. 15, 84-85, 146
Diehl, R. L. 87
Dolan, R. 267
Dotow, D. 270
Downey, A. 60
Drayson, Z. 35, 76-77, 90, 224, 302
Dretske, F. 149, 201
Dreyfus, H. L. 293
Duchliński, P. 11, 14
Dubs, A. 45, 116

E
Eckstein, M. 261
Egan, F. 60
Egbert, M. 15
Elkin, E. 109-110, 119, 122, 298
Elqayam, S. 30, 92, 120-125, 127-128, 199, 282, 303
Enoch, D. 265-266
Ernst, M. O. 88
Evans, J. S. 30, 91, 120-125, 127-128, 199, 282, 303
Evans, M. 169
Ey, H. 25
Eyer, J. 184

F
Fauquet, J. 136
Feldman, H. 42, 165
Felleman, D. 43, 166
Festinger, L. 266
Firth, R. 18
Fishburn, P. C. 249
Fiske, S. T. 61
Fitelson, B. 108
FitzGerald, T. 267
FitzPatrick, W. J. 16
Fortier, M. 26, 168-169, 175, 178-179, 240
Friedman, D. A. 26, 168-169, 175, 178-178, 240
Friston, K. J. 15, 25-26, 31, 41-44, 46, 62, 64, 69, 87, 92, 94, 107, 116, 131-146, 148,

150-152, 158-159, 161-162, 164-172, 174-175, 178-184, 193, 204-206, 210-211, 213, 215, 226, 238-248, 251-255, 257-258, 260-261, 263-264, 268, 273, 275, 277, 283, 286, 292-298, 303-304
Froese, T. 143-149, 151-152, 155, 157-158, 161, 163-164

G
Galanter, E. 81
Gallagher, S. 83, 252
Garson, J. 126, 181-182, 184, 197, 204, 217-218
Gawande, A. 283
Geisler, W. S. 87
Gibson, J. J. 76, 80-82, 86, 88, 214-215, 276, 294
Giere, R. 17
Glanville, R. 150
Glennan, S. 168, 218
Gładziejewski, P. 33, 40, 49-50, 52, 57, 59-60, 62, 64, 71-74, 104, 118, 132, 152, 202, 208, 212-213, 216, 219, 224, 228, 232, 287, 289, 297
Godfrey-Smith, P. 144-145, 152-154, 156-157, 180, 183-184, 187-188, 206, 244, 287
Goel, A. K. 168
Goldman, A. 70, 78
Goodman, N. D. 104
Gordon, N. 107
Gottlieb, J. 262
Graf, E. W. 88
Graves, A. 256
Green, D. M. 103
Greenman, J. 149
Gregory, R. L. 47, 53, 60, 74, 76
Griffiths, P. E. 153-154
Griffiths, T. L. 57, 103-104
Griffith, T. W. 168
Grobler, A. 91
Guo, X. 262

H
Hahn, U. 120
Haidt, J. 15
Harman, G. 118
Harkness, D. L. 25, 93, 104, 111, 232
Harrison, C. W. 46
Harrison, L. 132, 164-165, 167, 247
Hartmann, N. 14
Hartmann, S. 109-110, 118-119, 122, 298, 306
Haugeland, J. 274
Heinrich, D. 55
Heller, M. 147
Helmholtz, H. von 48-49, 87, 90-94, 160, 294
Hempel, C. 117, 140, 187
Henderson, D. 14-15
Heras-Escribano, M. 289
Hesp, C. 61, 237, 247, 252, 255, 297
Hildreth, E. C. 231
Hintikka, J. 97
Hinton, G. E. 39, 118
Hipólito, I. 103
Hohwy, J. 6, 24-25, 36, 38-40, 44, 48-50, 52, 56-57, 59-66, 68, 70-71, 74-75, 77, 87, 92-94, 102, 104, 107, 109, 117-118, 132, 139, 141, 152, 158-164, 166-167, 169, 173-174, 184, 193, 202, 208, 215-216, 228, 232, 242-243, 245-247, 253, 255, 260, 284, 287, 295, 298
Hookway, C. 53
Hopfield, J. J. 169
Höffe, O. 51
Hume, D. 11-12, 15
Husserl, E. 13
Hurley, S. 81
Hutchins, E. 62, 284
Hutto, D. 84, 86, 143-145, 149-150, 152, 157, 164

I
Ieropoulos, I. 149
Illari, P. 113, 190, 218
Ingarden, R. 12, 14-15
Isaac, A. 154

J
Jasrasaria, D. 235
Jäkel, F. 103, 114

Jensen, F. V. 100
Jensen, K. 20
Juechems, K. 262
Juś, M. 191, 193

K

Kaaronen, R. O. 266
Kahneman, D. 117, 251, 259-260, 266, 270, 282
Kalaska, J. F. 81
Kalinowski, J. 15
Kameyama, M. 261
Kant, I. 5, 13, 29, 49-57, 113, 169
Kaplan, D. M. 62, 105, 168, 199, 218, 229, 232
Kauffman, S. 164, 172
Kawalec, P. 4, 34, 91, 97, 100, 108-109, 111
Keijzer, F. 145
Kemp, C. 103
Kelsen, H. 15
Kelso, J. S. 295
Keshava, A. 25, 93, 104, 111, 232
Keynes, J. M. 97
Kiebel, S. J. 25, 44, 116, 132, 139, 165-167, 183, 211, 246-247, 257
Kiefer, A. 36, 39-40, 59-60, 71, 78, 118-119, 152, 202, 208, 215, 255, 287, 295
Kilner, J. 69, 132, 164-165, 167, 247
Kim, J. 18
Kirchhoff, M. 61, 79-80, 82, 87, 139-149, 151-152, 155, 157-158, 161, 163-164, 239, 247, 294-297
Kirmayer, L. J. 289
Kitcher, P. 18
Kiverstein, J. 25, 60, 79, 82, 86, 137, 143, 237, 241, 247, 292-294
Klawiter, A. 81
Klein, C. 132, 174, 210, 237, 242-243, 247, 283, 286
Knill, D. 111
Kolchinsky, A. 154-155, 200
Kołmogorow, A. 95-96, 108, 136
Korbak, T. 33, 142, 144, 171, 256, 261
Korsgaard, Ch. 11, 12, 14, 264-265
Kretzmer, E. R. 46
Kripke, S. 14
Kroes, P. 185
Krohs, H. 185
Krokos, J. 33, 266
Kwisthout, J. 43, 69, 101, 104, 111
Kyburg, H. 97-98

L

Latawiec, A. 147
Lau, J. 60
Laudan, L. 85, 109
Laughlin, S. B. 45, 116
Lee, G. 59, 79
Lee, T. S. 46
Leszczyński, D. 13-14, 52, 54
Lewis, D. 16
Levins, R. 169
Lieder, F. 104
Little, M. 16
Longo, G. 172
Lubański, M. 147
Luciw, M. 254
Lukac, M. 261
Lyon, P. 145, 196
Lyu, S. 87

M

Ma, W. J. 215
Machamer, P. 114, 219-220, 226, 269
Maciejczak, M. 12, 25, 50-51
MacKay, D. M. 47
Maher, Ch. 150
Makhoul, J. 45
Malaterre, C. 169
Markow, A. 5-6, 9, 31, 61-63, 100-103, 136, 139-142, 146-147, 159, 166-167, 246-247, 253, 257, 261, 303-304
Marr, D. 11, 30, 45, 91, 105-106, 111, 114-115, 123-125, 127-128, 174, 178-179, 216, 229-232, 262-263, 268, 271, 274, 289, 305
Mattout, J. 69, 165, 247
Maturana, H. 80, 84, 143, 203
Maull, N. 279

Maynard-Smith, J. 153
Mayr, E. 147
McAllister, J. W. 67-68
McDowell, J. 76
McGregor, S. 297
McKay, R. T. 70
Melhuish, C. 149
Metzinger, T. 24, 41, 92, 132
Miller, G. A. 81
Miller, M. 237, 241, 293-294
Miller, P. 293
Millikan, R. G. 6, 15, 18, 24, 121, 125-126, 168, 189-190, 193-198, 200-201, 203, 205-207, 270
Miłkowski, M. 33-34, 51, 55, 59-60, 71-72, 74, 83, 85, 91, 110, 113-114, 145, 147, 149-150, 171, 213, 228, 231, 269-270, 279, 281, 284, 289, 298
Mitchell, S. 279
Montague, R. 210, 215, 252-253, 257, 263
Montero, B. 17
Montévil, M. 172
Moń, R. 14
Moore, G. E. 16, 75
Moreno, A. 227
Morgenstern, O. 249, 251
Mortimer, H. 29, 95, 97, 108
Morville, T 247
Mossio, M. 269
Moutoussis, M. 252
Muders, S. 125
Muldoon, R. 15
Mumford, D. 46, 237
Müller, A. 85
Müller, K. 85
Myin, E. 84, 86, 143-145, 149-150, 152, 157, 164

N
Nair, V. 39
Nash, J. 20, 251
Neander, K. 189
Neisser, U. 47, 81
Nersessian, N. J. 168
Ng, A. Y. 38
Nie, L. 270
Noë, A. 84
Norman, D. A. 83, 89, 214, 232, 271-272, 276
Nowakowski, P. 33, 169

O
Oaksford, M. 103-104, 117, 120
Obidziński, M. 33
O'Brien, G. 71
Opie, J. 71
Ofner, A. 261
Oliver, B. 46
O'Regan, J. K. 84
Orlandi, N. 5, 46-47, 49, 53-54, 57, 59-60, 79-80, 82, 87-89, 99, 110, 173, 247
Oudeyer, P. Y. 262

P
Papineau, D. 17
Park, H. J. 166
Parr, T. 107
Pattee, H. H. 268, 274
Pearl, J. 61, 94, 101
Penny, W. 247
Perry, J. 270
Pezzulo, G. 193, 240, 247, 286
Pfeifer, R. 80
Piccinini, G. 172, 220, 232
Pickering, M. J. 38
Piekarski, M. 14, 34, 58, 121, 211, 213-214, 225, 286, 289-291
Piłat, R. 25
Platon 11
Poczobut, R. 147. 205
Poggio, T. 231
Pouget, A. 111
Povich, M. 279
Półtawski, A. 25
Pribram, K. 81
Protevi, J. 83
Pyzer-Knapp, E. 235

R
Ramsey, F. 98
Ramsey, W. M. 72, 74, 147-148
Ramstead, M. J. D. 33, 62, 134, 139, 141, 148, 159, 164, 167, 239, 246, 260, 264, 289, 294-297
Rao, R. P. 36, 43, 45-46, 107, 116, 183, 236
Ravenscroft, I. 221
Rączaszek-Leonardi, J. 269
Rescorla, M. 94, 99
Ridge, M. 16
Riese, K. 73
Rietveld, E. 79, 82, 86, 137, 139, 143, 214, 237, 241, 247, 255, 291-294
Rigoli, F. 240, 286
Robertson, I. 79-80, 87, 164
Rohde, M. 15
Roepstorff, A. 87, 107, 215
Rosati, C. S. 264
Rosenblueth, A. 181, 183
Rosch, E. 80, 83-84, 86
Rosvall, M. 156
Rubinov, M. 102

S
Sady, W. 85
Salmon, W. C. 14, 270
Savage, L. J. 98, 250
Scheler, M. 14
Schetz, A. 4, 34, 82-83
Schwartenbeck, P. 63, 175, 210, 238, 241, 246-247, 253-254, 257-258
Searle, J. 54, 158
Sejnowski, T. J. 118
Seth, A. K. 25, 40, 55-56, 117, 139, 149, 174, 184, 197, 286, 298
Seriès, P. 104, 106, 111, 114
Shagrir, O. 71, 105, 229-230, 271
Shams, L. 215
Shannon, E. A. 5, 131, 134-137, 144, 149, 152-154, 156, 161
Shea, H. 60, 71-72, 74, 156-157
Shi, Y. Q. 36
Silverman, D. 83
Simon, H. 104
Simoncelli, E. P. 87
Sims, A. 98, 169, 236, 239, 247, 278
Skyrms, B. 144, 156-157
Smith, K. C. 153-154
Sober, E. 194
Solopchuk, O. 248
Sontuoso, A. 15
Sporns, O. 102
Spratling, M. W. 36, 45, 109, 115-116
Sprevak, M. 60
Srinivasan, M. V. 45, 116
Stanford, P. K. 119
Stephan, K. E. 15, 26, 43, 119, 132-134, 136-137, 146, 165, 172, 193
Sterelny, K. 144, 152-154
Sterling, P. 184
Stern, R. 53
Stier, M. 125
Stober, S. 261
Summerfield, Ch. 262
Sun, H. 36
Susskind, J. 39
Swanson, L. R . 50-52, 54, 57
Swets, J. A. 103

Ś
Ślaga, S. 147

T
Tabery, J. 279
Taylor, S. E. 61
Tenenbaum, J. B. 103, 114
Thagard, P. 19, 260
Thompson, E. 80, 83-84, 86, 146, 150
Thornton, C. 111, 136, 148, 152, 166, 239
Tomasello, M. 15, 20
Tooby, J. 111
Turing, A. M. 111, 175
Tversky, A. 117, 251, 259-260, 266, 270, 282
Tyszka, T. 96, 98, 111, 249-250

U
Umerez, J. 269
Ullman, S. 231
Uribe, R. 80

Q
Quine, W. V. O. 18

W
Wachowski, W. 15, 33, 62, 121, 214, 289-291
Waleszczyński, A. 14, 33
Walker, R. C. S. 53
Wallace, D. 69
Ward, D. 83
Warren, W. H. 81
Watson, S. 183
Weaver, W. 136
Weiskopf, D. A. 231
Whiting, D. 14
Wiener, N. 181, 183
Wiese, W. 24, 41, 59, 70, 92, 132, 152
Williams, D. 44, 113, 174, 266, 280, 282, 286
Williamson, J. 113
Wright, C. 26, 168, 170-173, 247, 253
Wright, L. 6, 153, 182, 186-189
Woodward, J. 67
Woleński, J. 91
Wolpert, D. H. 154-155, 200
Wouters, A. G. 6, 186-187, 190-194, 200

V
Varela, F. 80, 83-86, 141-143, 150, 203
Van der Cruys, S. 243
van Duijn, M. 145, 196
Van Es, T. 265, 298
van Essen, D. 43, 166
van Fraassen, C. 344
Van Gelder, T. 26, 168
van Rooij, I. 43, 69, 101, 104, 111, 297
Veissière, S. P. L. 110, 259-260, 289
Velleman, J. D. 264
Villalobos, M. 83, 85, 142, 150
Verbeurgt, K. 260
Vogel, J. 65
von Mises, L. 97
von Neumann, J. 249, 251
von Wahlert, G. 191

Y
Ylikoski, P. 14, 114

Z
Zahavi, D. 30, 35, 49-50, 57-62, 64-65, 68, 70, 74-79, 89-90, 302
Zaleśkiewicz, T. 96, 111
Zednik, C. 24, 29, 51, 72, 103-105, 114, 219, 269

Indeks rzeczy

A

adaptacja 134, 137, 143, 146, 172, 198, 283, 296
afordancja 7, 82, 83, 86, 143, 215, 255, 272, 276-277, 288-295
  epistemiczna 240, 294
aksjomat 100, 108, 173, 249-251
  Kołmogorowa 95-96
Analiza Racjonalna 103-104, 106-107, 216
algorytm 5, 22, 30, 40, 45, 57, 66, 91, 104-107, 111, 115-116, 119, 124, 128, 174, 175, 229-230, 236, 255-257, 261, 278, 287, 302
allostaza 184, 243
antycypacja 209, 214
antynaturalizm 15-16
antyrealizm 30, 35, 49, 58-59, 61, 64, 70, 74, 77-79, 90, 92, 108, 119, 123, 127-129, 200, 297-298, 302-303
antyreprezentacjonizm 59, 82-84, 86-87, 89, 162-164, 174, 294
aproksymacja 93, 115, 118, 216, 260
argument
  konstytutywny 264-266
  transcendentalny 30, 53-57, 170-172, 174, 302
artefakt 146, 149, 170, 183, 187-188, 194-195, 285, 291
autonomia 83-84, 146, 149-151, 155, 166, 172, 203-204, 208-209, 264-265
autopojeza 84-95, 140-143, 146-151, 155, 164, 203-204
analiza przez syntezę 52, 90

B

bayesizm
  normatywny 120, 122
  oświecony 5, 91, 112, 115, 123, 128, 302
  preskryptywny 91, 121, 303

C

ciekawość 241, 293
cybernetyka 85, 137, 148-149, 151, 181-183
dedukcja transcendentalna 29, 35, 51, 55, 302
deskryptywizm 30, 91, 120-122, 124, 128, 303

D

długoterminowa minimalizacja 32, 37, 233, 238, 244-245, 265, 277, 280, 284, 286
długoterminowe średnie zaskoczenie 134, 167, 238, 242
dopamina 267
dowód dla modelu 252
dynamicyzm 86, 164, 269, 292
dysfunkcja 15, 31, 125-126, 145, 153, 177, 185-186, 188-189, 193, 195-199, 201, 203, 205, 212, 217-218, 304
dysjunktywizm 76
dywergencja Kullbacka-Leiblera (K-L) 159, 202, 254, 267, 277, 292
dylemat eksploracji i eksploatacji 32, 233, 263, 293

E

eksternalizm 62, 213, 287
enaktywizm
  klasyczny (autopojetyczny) 80, 84
  obliczeniowy 295
  partycypacyjno-sensotwórczy 84
  radykalny 84, 86,
  sensomotoryczny 84
energia swobodna (wariacyjna) 26, 116, 131, 133-134, 138-141, 147, 152, 158-159, 161, 164, 166-167, 169-175, 178-180, 184, 236, 239, 242, 246, 252-254, 258, 283, 292, 295-296, 304

entropia 134, 136-137, 139, 154, 156, 166, 171, 173, 175, 178-180, 183, 185, 292
ergodyczność 134, 142, 147, 172, 178, 247
etiologia (etiologiczny) 187-189, 193-194, 196-198, 200, 203, 206-208, 217, 220
(neuro)etologia 206, 247
ewolucja (teoria e.) 15, 111, 120, 153, 160, 169, 172, 182, 191, 196, 198-199, 211, 239

F
fenotyp 136-137, 153, 182, 252, 295-296
fundamentalizm bayesowski 91, 112, 114, 128, 302
funkcja
biologiczna 18, 177, 192, 200, 202, 275
eskplanacyjna 72, 161-162, 185
etiologiczna 194
heurytyczna 85
konstytutywna 13
normatywna 15, 18, 23-24, 31-32, 126, 138, 146, 175, 177, 184-186, 190, 196, 199-200, 202, 204-206, 212, 217-218, 221, 225-228, 234, 253, 274, 280, 304, 306
pochodna 194-195
predykcyjna 31-32, 177, 200, 212, 216, 225, 227, 234, 267, 275, 289, 304-306
przyczynowa 27, 32, 72, 177, 205, 282, 285, 290, 304
reprezentowania (reprezentacyjna) 6, 31, 177, 200-202, 208-209, 212-213, 227, 304
systemowa 186, 188
użyteczności 249-250
właściwa 6, 194-197, 201

G
genotyp 137, 182

H
heurystyka
badawcza 31, 131, 174-175, 178, 237, 259, 297, 304
umysłowej dostępności 259, 270
poznawcza 259, 270
reprezentatywności 259, 270
hierarchiczność 5, 22, 25, 36, 40-41, 43-44, 46, 55, 92-94, 98, 104, 106-107, 109, 117-118, 132, 166-167, 225, 227, 244-245, 247-248, 253, 258, 260-261, 274, 279, 284
hipoteza
antyreprezentacjonistyczna 162-163
eksplanacyjna 23, 128
deskryptywna 298
epistemiczna 23, 26, 31-32
kontrfaktyczna 286
normatywności rozproszonej 7, 33, 234, 288-291, 299
ontyczna 22, 26-27, 29, 31-32, 128, 282
percepcyjna 47-48, 50, 66, 88-89
reprezentacjonistyczna 162-163
homeostaza 6, 15, 31, 61, 121, 125, 132-134, 136-139, 142, 146, 148, 159-160, 165-166, 168, 170-171, 173, 178, 180-185, 191-192, 197, 206, 240-241, 246, 254, 278, 296, 303-304
homeostat 181

I
idealizacja 13, 21, 81, 108-109, 173, 237, 279, 283, 297
inferencjalizm 58, 60, 87, 89-90, 294
informacja
biologiczna 144, 152-153
semantyczna 131, 142-145, 151-155, 157, 161, 175, 180, 295
syntaktyczna (Shannona) 5, 131, 134, 136, 144, 149, 151-156, 158, 161, 180
wzajemna 145
instrumentalizm 128, 163-164, 200, 298
bayesowski 112-114
intencjonalność 13, 21-22, 84, 144-145, 154-155, 185, 195, 222-223, 225, 293
internalizm 5, 30, 35, 58-62, 70, 78-79, 90, 101, 103, 139-140, 207-209, 213, 263, 302
integracja 7, 32-33, 233, 247, 278-281, 286, 298-299, 305

introcepcja 184
inżynieria odwrotna 5, 30, 35, 51-53, 56-57, 90, 113-114, 302
irracjonalność 124, 265-266

K
koc Markowa 5, 6, 31, 60-63, 100-103, 136, 139-142, 146-147, 159, 166-167, 246-247, 253, 257, 261, 303-304
komponent
konstytutywny 22, 32, 232, 234, 280, 288, 299, 305
mechanizmu 4-5, 24, 72, 105-106, 111, 168, 172-173, 219, 229-232, 234, 263, 269, 272, 284, 288, 307
kognitywizm 6, 27, 131, 142-143, 151-152, 158, 161-164, 171, 173, 303
konstruktywizm 49, 53, 59, 76-77, 79, 87-89, 294
konstytucja 13, 51-52, 115, 147, 263, 274

L
loteria 249-250

Ł
łańcuch Markowa 101, 136, 139

M
mapowanie 168, 219, 229-230
mechanicyzm 14, 21, 161, 164, 167-168, 170-173, 218, 227-228, 231, 269, 271
mechanizm
allostatyczny 184
biologiczny 24, 185, 207, 303
homeostatyczny 31, 180-184, 192, 206, 296, 304
konstytutywny 14, 288, 290
minimalizacji energii swobodnej 31, 131, 161, 175, 206, 303-304
motywacyjny 275
normatywny 6-7, 13-14, 17-18, 21, 23-27, 29-34, 38, 45, 62, 64, 92, 104-105, 116, 126, 137-138, 145, 165, 167, 172, 174-175, 177, 183, 185-186, 197, 206-207, 210-212, 215, 217-218, 221, 227-229, 231-234, 237, 243, 255, 261-262, 267-268, 271, 274, 278-282, 285-291, 299, 301, 304-307
predykcyjny 6, 13, 21-22, 27, 31-33, 131, 174-175, 177-178, 184, 186, 190, 192-193, 195-197, 200, 202, 207-208, 211-212, 216-218, 220-221, 225, 231-234, 236, 242-243, 245, 261, 278, 280-281, 284-292, 298-299, 301, 303-306
regulacyjny 148, 240, 283
reprezentacyjny 207-209, 212, 280
teleologiczny 206
termoregulacji 145, 182
moc eksplanacyjna 106, 111, 133, 147, 158, 161, 172-173, 185, 187, 218, 232, 278, 281-282, 305
model
bayesowski 5, 12, 25, 30, 39, 42, 45, 57, 91-92, 100, 103, 106, 108-112, 114-117, 119-120, 122-123, 128, 132, 167, 185, 196, 200, 216, 229, 251-253, 265, 282, 297-298, 302-303
dyskryminacyjny 38
dynamiczny 168
eksplanacyjny 5, 42, 45, 57, 91, 120, 164, 185, 196, 216, 229-230, 296, 302
funkcjonalny 168
generatywny 5, 9, 25, 35-41, 43-45, 52, 55, 59, 61-63, 66, 68, 70, 74, 78-79, 89-90, 92-94, 97-98, 101, 107, 109, 111, 116-118, 127, 132, 138-141, 147, 152, 159, 163, 167, 174, 178, 183-184, 192-193, 196-197, 200, 202-203, 210-214, 216-217, 225, 228, 234, 238, 242-247, 266-267, 274-275, 284, 286-289, 291, 293-298, 302
generyczny 168
inwersyjny 166
mechanistyczny 6, 18, 21-22, 62, 90, 168, 216, 217, 244, 285-286, 290
obliczeniowy 40, 118, 122
pojęciowy 169

probabilistyczny 39, 103, 232, 282, 286
przetwarzania predykcyjnego 5, 12, 24, 91-92, 115, 120, 128, 174, 178, 197, 236, 251, 274, 282, 285-286, 303
replikacyjny 114
statystyczny 38, 39
świata 25, 38, 42, 47, 59, 70, 74, 83, 92-93, 117, 134, 148-149, 160, 212-213, 239, 244, 252, 258, 263, 268, 273, 277, 281, 292, 294
monizm eksplanacyjny 33, 234, 280, 282-283, 298, 305
motywacja 6, 33, 209, 215, 233, 242-245, 258, 261-264, 266-268, 272-275, 278, 280, 283-286, 306

N

nagroda 63, 198, 215, 244, 249, 253-254, 256-258, 262, 285
nastawienie 6, 13, 21, 86-87, 119, 221-222, 262, 266-267, 272-273, 275, 278
neuroreprezentacjonizm 49, 58-59, 62, 74
niepewność 6, 32, 42, 92, 99, 101, 119, 134-136, 140, 145-156, 166, 175, 192, 210-211, 213, 214-215, 217, 221, 233, 235-255, 257-259, 261, 267, 272-273, 277, 283-286, 294, 296, 306
nonkognitywizm 6, 27, 131, 142-143, 151-152, 158, 161-164, 171, 173, 303
normatywność
a moralność 11, 14-16
biologiczna 17-18, 149
eksplanacyjna 22, 24, 26, 301, 306
FEP (zasady energii swobodnej) 165, 175, 177-178, 185, 253
funkcji 31, 126, 145, 153, 190, 200, 202, 206, 217, 227, 234, 306
homeostazy 6, 121, 125, 148, 180, 184, 206, 241, 278, 296
konstytutywna 22, 290
mechanizmu 27, 32, 175, 177, 184-186, 197, 199-200, 202-203, 210, 211, 216-217, 227, 231, 234, 236, 242, 282, 287, 298, 304
modelu 5, 123, 149, 202, 215, 251, 282, 296,
oceniająca 19, 22, 301, 303
predykcji 21, 211, 213-214, 274
przetwarzania predykcyjnego 27, 116, 120, 209, 282, 304
reprezentacji 6, 208-209, 225
struktury 290
środowiska 86, 281
rozproszona 7, 33, 234, 288-291, 299, 307
usytuowana 291
wewnętrzna 22, 207, 208, 301, 303
zewnętrzna 207
normowanie 12, 202, 208, 214, 287
nowość 235, 241, 248, 254-255, 258-259, 262, 264, 267, 283

O

organicyzm 164, 170-172
ograniczenie
3M 168, 219
biologiczne 110, 120, 263,
chemiczne 275
eksplanacyjne 280
empiryczne 270
ewolucyjne 110, 266
fizjologiczne 275
fizyczne 230-231, 263, 272, 274-275, 281, 284
konstytutywne 282
konwencjonalne 284
kulturowe 231, 263, 272, 274, 281
logiczne 272, 274, 281, 284
metodologiczne 269, 281, 284
mechanizmów 233-234, 264, 273, 280-281, 287
normatywne 231
obliczeniowe 110, 266
semantyczne 271-272, 274
społeczne 231, 263
środowiskowe 38, 234, 243, 272-273, 275-276, 283, 291, 299, 305
wewnętrzne 263-264
zewnętrzne 263-264, 267-268, 285-286

P
paradoks
Allais'ego 251
Ellsberga 251
pełne twierdzenie klasowe 252, 254
pluralizm eksplanacyjny 237, 281
podejmowanie decyzji 6, 37, 81, 98, 112, 121-123, 132, 212, 215, 228, 244, 248-251, 253-255, 259, 278, 280, 296
podobieństwo strukturalne 60, 71-74, 79
pośredniość/bezpośredniość percepcji 5, 25, 35, 74-77, 89-90, 302
poziom
algorytmiczny 30, 31, 91, 105, 111, 114-115, 123-125, 128, 174
implementacji 30, 31, 91, 105-106, 111, 114, 124-125, 128, 302
konstytucji 51
mechanistyczy 112, 114, 128, 288, 302
obliczeniowy 30, 31, 105-107, 111-112, 114, 124, 128, 179, 229-230
opisu 5, 23
osobowy 44, 224-225, 288
subosobowy 6, 23, 44, 220, 224-225, 290
wyjaśniania 6, 112, 128, 220, 224-225, 243, 288, 290
praktyka
badawcza 218, 228, 281
ewaluacyjne 21-22
kulturowa 143, 291
normatywna 290-291
prawdopodobieństwo
pierwotne 95-96, 98-99, 101, 225
rozkład p. 30, 38-39, 42-43, 64, 91, 98-101, 103, 106, 109, 112, 117-118, 136, 152, 156, 159, 164, 202, 210, 213, 224, 243, 249, 254, 267-269, 284, 292, 294
warunkowe 95-96, 100-101
wspólne 38-39
wtórne 96, 100-101, 224, 267
względne 99
precyzja 5, 41-44, 68-69, 71, 73, 92, 98, 109, 118, 166-167, 243, 248, 261, 267, 287
preferencja 13, 99-100, 210, 215, 245, 249-251, 259, 263, 266-267, 286
preskryptywizm
epistemiczny 5, 92, 127-129, 185, 196, 199-200, 216, 298, 303
mocny 30, 92, 122-123, 127-128, 303
ontyczny 5-6, 32, 92, 127-129, 177-178, 185, 199, 216, 218, 232, 298, 303-305
słaby 30, 92, 122-123, 127-128, 303
problem ciemnego pokoju 6-7, 32-33, 174, 210-211, 214, 217, 233-234, 236-245, 254, 258, 262, 274, 277-278, 280, 282-283, 285, 293, 296, 298, 305-306
przyczyna konstytutywna 20-21, 23-24, 217, 221, 288-290
psychologia
ekologiczna 5-6, 80, 82, 86, 88, 90, 255, 270, 294
postaci 81
potoczna 13, 23, 221-225, 243

R
racja 15, 18, 221-223, 243, 247,
racjonalność
ekologiczna 120-121
ewolucyjna 120-121
instrumentalna 120-121
normatywna 91, 120-121
ograniczona 120
realizm
bayesowskiego PP 30, 66, 92, 123, 127, 129, 178, 216, 303, 305
epistemologiczny 78
etyczny 16
metafizyczny 78
naiwny 76
naukowy 58, 72
normatywny 17
regulator (t. o dobrym regulatorze) 148, 283
relacja
konstytutywna 165, 289, 294
motywacyjna 13, 23, 214, 275, 286-287, 290
normatywna 146, 214, 287, 291

reguła (twierdzenie) Bayesa 30, 39, 91-92, 94, 96-99, 101, 104, 106-107, 109, 111, 117-123, 127-129, 141, 196, 215-216, 275, 277, 291, 298, 303
reprezentacja
błędna 126, 153, 201-203, 209, 215, 217, 287
strukturalna (S-reprezentacja) 5, 30, 35, 59, 62, 70-74, 79, 90, 102, 109, 145, 213, 287, 295, 302
reprezentacjonizm 49, 58-61, 84, 86-87, 89-90, 147-149, 152, 162-163, 174, 292-293
rola
eksplanacyjna 30, 223, 226, 286
konstytutywna 17, 68
przyczynowa 21-24, 29, 153, 205, 218, 228, 232, 262, 273, 281, 284-285, 295, 301
rozkład
normalny (Gaussa) 42
prawdopodobieństwa 30, 38-43, 64, 91, 98-101, 106, 109, 112, 117-118, 136, 152, 156, 159, 202, 210, 213, 224, 243, 249, 254, 267-269, 274, 292, 294
rozpoznawanie wzorców 65-66, 68, 73, 116, 228, 232, 274
równowaga Nasha 20, 251
ryzyko (warunki ryzyka) 20, 37, 93, 203, 214, 235, 243, 249-250, 258, 277, 284

S
samoorganizacja 61, 83-85, 133, 136, 139, 141, 143, 146-148, 151, 155, 166-167, 169, 203, 253, 295
selekcja naturalna 145, 153, 155, 160, 165, 167, 169, 172, 187-189, 198-199, 201, 211
sieć bayesowska (przekonań) 9, 25, 61, 78, 91, 94, 100-101, 105, 109, 111, 118, 165, 212, 238, 248, 258, 260-261, 265, 275, 286
skierowany graf acykliczny (DAG) 9, 100, 102
sprzężenie zwrotne 46, 63, 116, 181-182
stabilność 31, 177, 180, 182, 184, 204-205, 212, 226, 234, 274, 284, 288, 304
statystyka scen naturalnych 87-89, 110
szkic mechanizmu 104, 172, 232, 278

T
teleosemantyka 6, 21, 31, 155, 194, 198, 200, 202, 205, 207, 218, 287, 306
druga zasada termodynamika 134, 151, 165, 226
transcendentalizm 13, 17, 52
trudny problem treści 143-145, 149

U
unifikacja 25, 168, 179
uczenie
maszynowe 61, 65, 104, 234, 255, 257-258, 260-261, 263
nadzorowane 249, 255-258, 261
nienadzorowane 255-256, 261
przez wzmocnienie 255, 257-258, 261
uwaga 25, 42, 68, 98, 165, 238

W
wartość
epistemiczna 165, 236, 248, 254, 263, 284
logiczna (prawdy i fałszu) 31-32, 177, 208-209, 212-213, 217, 304
nieepistemiczna 236
oczekiwana 248-250, 252
pragmatyczna 248
warunki
logiczne 14, 20
transcendentalne 13, 171
ważenie precyzji 41, 44, 68-69
weredyczność 76-77, 149, 280
wiarygodność/wierzytelność (*self-evidence*)
modelu 63, 68, 79, 215, 248, 252-255, 257, 260, 264-266, 270, 273, 275, 277, 285, 287, 294, 296, 298
przekonań 252, 254, 259
własność konstytutywna 22-23, 264
wnioskowanie
abdukcyjne (do najlepszego wyjaśnienia) 30, 35, 50, 56, 60, 78, 90, 117, 140, 159, 171, 212, 283, 286, 302
aktywne 6-7, 20, 32, 64-65, 69-70, 118, 140-142, 146-147, 149, 159, 167, 178,

210, 233-234, 241, 244-249, 251-255, 258-267, 273, 277-278, 280, 283, 292, 295-299, 305
bayesowskie 55, 93-94, 99, 110-112, 115, 117-119, 127, 140-141, 216, 253, 283, 292, 294
enaktywne 7, 118, 138, 255, 265, 295
indukcyjne 115, 117
materialne 275
odwrotne 56
percepcyjne 63-64, 76, 89, 119, 138, 159, 165, 167, 246
probabilistyczne 45, 94, 103-104, 106-107, 112, 117-119, 217, 252
redukcyjne 56-57
statystyczne 55, 93-94
„tak jakby" 297
wyjaśnienie
dynamiczne 269
etiologiczne 207, 220
funkcjonalne 57, 191, 193
horyzontalne 224
konstytutywne 14, 220
mechanistyczne 6, 104, 113-114, 218, 225, 227, 269
przyczynowe 71, 74, 79, 200, 202, 221
wertykalne 224
wzór/wzorzec
działania/zachowania 15, 20, 112, 175, 262, 287-288
pobudzeń 62
rzeczywiste/realne 30, 35, 59, 65-66, 70-71, 73, 79, 108-110, 213, 245, 274, 287, 289
sensomotoryczne 83
sensoryczne 64, 79
statystyczny 35, 63, 69-70, 73, 79, 90, 256, 302
współaktywacji 72

Z

zasada
działania (*action selection/policy selection*) 241, 243, 247-248, 252, 254-255, 257-259, 263, 267, 283, 295, 306
energii swobodnej (FEP) 5, 6, 25-26, 31-32, 45-46, 116, 131-133, 137, 139-140, 142-143, 145-156, 148-150, 152, 158-161, 163-175, 177-180, 183-185, 206, 210, 227, 233, 236-244, 249, 251-254, 257-258, 260-261, 278, 282-283, 286, 292, 295-298, 303-305
epistemiczna 247
maksymalizacji wartości oczekiwanej 249
metodologiczna 269-270, 279
normatywna 26-27, 165, 168, 175, 178-179, 246, 282, 301, 303
pierwsza (*a priori*) 171, 173-175
(stacjonarnego działania) Hamiltona 169
tautologiczna 165
uaktualniania prawdopodobieństwa 109
warunkowania 97, 99
zaskoczenie 23, 42-43, 134, 137, 140-142, 145-146, 152, 158-159, 166-167, 171, 178, 210, 214-215, 236, 238-239, 241-242, 245, 247-248, 253, 255, 258, 277, 283, 292, 297-298

Table of contents

LIST OF FIGURES AND TABLES 9

INTRODUCTION 11

PREFACE 29

CHAPTER I. THE CONCEPT OF PREDICTIVE PROCESSING 35
- Introduction 35
- 1.1. General characteristics and basic concepts 36
  - 1.1.1. Prediction 37
  - 1.1.2. Generative model 38
  - 1.1.3. Precision 41
  - 1.1.4. Hierarchy 43
- 1.2. Possible predictive processing algorithms and philosophical inspiration 45
- 1.3. Kant and predictive processing 50
- 1.4. Critique of predictive processing 57
  - 1.4.1. Conservative predictive processing 59
    - 1.4.1.1. Internalism and Markov blankets 61
    - 1.4.1.2. Skeptical problem 63
    - 1.4.1.3. The real patterns in predictive processing 65
    - 1.4.1.4. S-representations and real patterns 71
  - 1.4.2. Indirectness of perception in predictive processing 74
- 1.5. Radical predictive processing 79
  - 1.5.1. Ecological psychology 80
  - 1.5.2. Enactivism 83
  - 1.5.3. Ecological psychology, enactivism and predictive processing 86
  - 1.5.4. Nico Orlandi's proposal 87
- Summary 90

CHAPTER II. PREDICTIVE PROCESSING AS BAYESIAN EXPLANATORY MODEL 91
- Introduction 91
- 2.1. Preliminary remarks 92
- 2.2. Basic concepts 94
- 2.3. Bayesian brain thesis 103
- 2.4. Levels of analysis and predictive processing 105
- 2.5. Problems with Bayesian models 108
  - 2.5.1. Explanatory problems of Bayesianism 110
  - 2.5.2. Reverse engineering in predictive processing 113
- 2.6. Predictive processing as enlightened Bayesianism 115
  - 2.6.1. Arguments for the explanatory nature of Bayesianism 115
  - 2.6.2. Normativity of the Bayesian model of predictive processing 120
    - 2.6.2.1. Normativity and levels of description 123
    - 2.6.2.2. Epistemic and ontic prescriptivism 127
- Summary 128

CHAPTER III. FREE ENERGY PRINCIPLE IN PREDICTIVE PROCESSING
- Introduction 131
- 3.1. Preliminary remarks 132
- 3.2. The concept of the free energy principle 133
  - 3.2.1. The concept of Shannon's information 134
  - 3.2.2. Homeostasis in free energy principle 136
- 3.3. Markov blankets and predictive processing 139

3.4. Free energy and the continuity of life and mind 142
3.4.1. Enactive approach 143
3.4.2. Critique of the enactive approach 147
3.4.3. Cognitivist Approach: Arguments 151
3.4.4. The Cognitive Approach: Hohwy's attitude 158
3.4.5. The noncognitivist approach versus the cognitivist approach 161
3.5. Explanatory difficulties of free energy principle 164
3.6. Application of the free energy principle in predictive processing 173
Summary 175

CHAPTER IV. NORMATIVE FUNCTIONS AND MECHANISMS IN THE CONTEXT OF PREDICTIVE PROCESSING 177
Introduction 177
4.1. Preliminary remarks 178
4.2. The normativity of homeostasis 180
4.3. The concept of a function 185
4.3.1. Functions according to Wright and Cummins 186
4.3.2. Functions in Wouters' approach 190
4.3.3. Functions in Millikan's approach 194
4.3.3.1. Proper functions and predictive mechanisms 195
4.3.3.2. Critique of the teleosemantic approach 198
4.3.4. Representational function 201
4.3.4.1. Functions in Bickhard's approach 203
4.3.4.2. The normativity of representations 208
4.3.4.3. Bickhard's objections on predictive processing 209
4.4. The normativity model for predictive processing 211
4.5. Predictive processing and the mechanistic model of scientific explanation 216
4.5.1. The normativity of predictive mechanisms and ontic prescriptivism 216
4.5.2. Mechanistic model of scientific explanation and normative mechanisms 218
4.5.2.1. Predictive mechanisms as causes and the subpersonal level of explanation 220
4.5.2.2. Possibility of mechanistic explanation of normativity 225
4.5.2.3. Explanation of normative mechanisms and constraints 229
Summary 232

CHAPTER V. NORMATIVE MECHANISMS AND ACTIONS IN PREDICTIVE PROCESSING 233
Introduction 233
5.1. Preliminary remarks 234
5.2. Dark room problem 237
5.2.1. Responses from supporters of predictive processing 239
5.2.2. Critique of predictive processing from the perspective of the dark room problem 242
5.3. Action under uncertainty 244
5.3.1. Active inference framework 246
5.3.2. Decision-making 248
5.3.2.1. Classical models 249
5.3.2.2. Critique of classical models from the perspective of active inference framework 251
5.3.2.3. Types of learning and active inference 255
5.3.3. Remarks on active inference framework 258
5.4. Normative predictive mechanisms and constraints 261
5.4.1. Normative mechanisms: motivation and attitude 262
5.4.2. Constraints 268
5.4.2.1. Importance of constraints for action 272
5.4.2.2. Constraints in predictive processing 273
5.4.2.3. Affordances - preliminary analysis 276

5.4.2.4. Limits of explanation in predictive processing and the postulate of integration 278
5.4.2.5. Constraints as conditions for integrating models 279
5.5. Explanation of normative mechanisms in the context of predictive processing 281
5.5.1. The dark room problem and normative mechanisms 283
5.5.2. The role of prediction 286
5.6. Normative mechanisms and the hypothesis of distributed normativity 288
5.7. Affordance-oriented predictions 291
5.8. Active inference as enactive inference 295
Summary 298

CONCLUSION 301

BIBLIOGRAPHY 309

SUMMARY 337

INDEX OF NAMES 341

INDEX OF ITEMS 349

Michał Piekarski (1984) – dr, adiunkt w Katedrze Teorii Poznania w Instytucie Filozofii UKSW w Warszawie. Jego zainteresowania badawcze dotyczą przetwarzania predykcyjnego, normatywności biologicznej i społecznej oraz wybranych zagadnień z zakresu filozofii kognitywistyki i epistemologii. Autor książki *Logika-Gramatyka-Pragmatyka. Ewolucja Wittgensteinowskiej koncepcji związku języka ze światem* (Wydawnictwo Naukowe UKSW, Warszawa 2014) oraz kilkudziesięciu artykułów opublikowanych w polskich i zagranicznych czasopismach naukowych.

Lektura książki Michała Piekarskiego pt. *Mechanizmy predykcyjne i ich normatywność* wywołała we mnie miłe zaskoczenie, gdyż sprawnie łączy w koherentną całość dwa, na pozór mocno rozbieżne, podejścia do zagadnienia predykcji i normatywności, tj. podejście klasyczne z podejściem naturalistycznym. Jest to zabieg trudny i wiele dotychczasowych prób dokonania takiego połączenia skończyło się niepowodzeniem. To pierwsze wiąże się z przekonaniem, że normatywność stanowi wewnętrzną cechę stanów, obiektów czy bytów, zaś to drugie zakłada, że normatywność ma charakter zewnętrzny, tzn. jest przypisywana systemowi przez obserwatora, zaś użycie do jej wyjaśnienia metody naukowej sprawia, że pojęcie normatywności staje się redundantne. Piekarski sprawnie i przekonująco argumentuje, że podejścia te można połączyć uzyskując realistyczno-naturalistyczne rozumienie normatywności, gdzie normatywność jest wewnętrzną cechą systemów ożywionych, stanowi element łańcucha wyjaśnienia kauzalnego, zaś pojęcie normatywności pełni w związku z tym, funkcję naukowo istotną (ma siłę eksplanacyjną, więc nie jest redundantne).

dr hab. Adriana Schetz, prof. US

Książka *Mechanizmy predykcyjne i ich normatywność* znacznie wykracza poza proste referowanie stanowisk prezentowanych w dotychczasowych dyskusjach i ujawnia wiele z ich ograniczeń i trudności. Jest to niewątpliwy walor książki i wyróżnik stylu dra Piekarskiego. Warto też podkreślić, że przedmiotem tego polemicznego odniesienia są najnowsze koncepcje w teorii umysłu i wiele ze szczegółowych spostrzeżeń Autora książki jest istotnym wkładem do dotychczasowych dyskusji. (...) Zasadnicze części książki ujawniają spójny odautorski dyskurs, który jest dość konsekwentnie przeprowadzony przez całą rozprawę i stanowi niewątpliwy jej walor. Czytelnik z ciekawością poddaje się kolejnym eksploracjom teoretycznym Autora i jest wciągnięty przez jego liczne, zróżnicowane tematycznie i - problematyzujące najnowsze koncepcje filozoficzne - oryginalne oraz drobiazgowe polemiki i dyskusje. W poruszaniu się po wielowątkowości i bogactwie tematycznym książki bardzo pomocne są też podsumowania poszczególnych rozdziałów oraz całości.

prof. dr hab. Paweł Kawalec

www.liberilibri.pl
ISBN 978-83-63487-44-7